Modelling leukemia in the mouse: novel strategies in genome engineering

Thesis

UNRESTRICTED

Modelling leukemia in the mouse:

novel strategies in genome engineering

Giuseppe Testa, M.D.

A thesis submitted in partial fulfilment of the requirements of the Open University for the degree of Doctor of Philosophy

April 2001

Sponsoring establishment
National Institute for Medical Research, London

Collaborating Establishment
European Molecular Biology Laboratory, Heidelberg

DATE OF SUBMISSION: 25 APRIL 2001
DATE OF AWARD: 4 APRIL 2002

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

Thesis abstract

The work of this thesis presents novel genome engineering strategies for assembling complex targeting vectors for functional studies in the mouse. With the objective of establishing a conditional mouse model of the human acute leukemia associated with the t(4;11)(q21;q23) translocation, an approach based on the Cre-loxP technology was chosen, and two mouse lines and one ES cell line were engineered starting from BACs.

For the *Af4* mouse line, a combinatorial mutant allele was designed, which supports Cre-driven interchromosomal translocation, can result in a knock-out or hypomorphic allele (also amenable to conditional gene repair strategies) and can report endogenous expression of the gene through β-galactosidase staining. The work utilised several variations of the ET cloning methodology and served to establish their feasibility for BAC engineering as a novel approach to the generation of complex mouse knock-in/knock-out targeting vectors. Particularly noteworthy was the application of ET technology to the direct subcloning of very large BAC fragments.

The *Mll* targeting construct constitutes the first example of a BAC based, very large knock-in vector (about 70 kb) which was successfully used to target the endogenous *Mll* locus by homologous recombination in mouse ES cells. The construct was engineered via ET recombination approaches. The major novelty consists in the targeting design which can simultaneously mutate two sites of the gene located very far from each other. Combined with the power of site specific recombinases (SSRs) technology, this opens the way to establishing complex, versatile mouse lines with only one round of ES cell targeting.

For the second mouse line, a BAC transgenesis approach was chosen to place Cre recombinase under the control of the *Ikaros* gene, in a new configuration which can yield

either constitutive or regulated Cre activity. Several BAC modifications were applied, and the usefulness of TnpI recombinase in BAC engineering was established.

Some of the results presented here appear in the following publications:

Muyrers, J. P., Zhang, Y., Testa, G., and Stewart, A. F. (1999). Rapid modification of bacterial artificial chromosomes by ET-recombination. Nucleic Acids Res *27*, 1555-1557

Testa, G. and Stewart, A.F. (2000). Creating a transloxation. Engineering interchromosomal translocations in the mouse. EMBO Rep. 1, 120-121

Zhang, Y., Muyrers, J. P., Testa, G., and Stewart, A. F. (2000). DNA cloning by homologous recombination in Escherichia coli. Nat Biotechnol *18*, 1314-1317.

This thesis is dedicated to my mother and my father.

Table of contents

List of figures

List of tables

List of abbreviations

AF4	The partner gene of *MLL* in the t(4;11)(q21;q23) translocation
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid leukemia
Ara	L(+)arabinose
ATP	Adenosine triphosphate
BAC	Bacterial Artificial Chromosome
bp	base pairs of DNA
BCR	Breakpoint Cluster Region
BSA	Bovine Serum Albumin
βGeok	Beta-galactosidase-Neomycin fusion
CBD	Calmodulin Binding Domain
CBP	Creb binding protein
Ci	Curie
Cre	The site specific recombinase encoded by bacteriophage P1
dATP	Deoxyadenosine triphosphate
dCTP	Deoxycytidine triphosphate
dGTP	Deoxyguanosine triphosphate
dTTP	Deoxythymidine triphosphate
DMSO	Dimethylsulfoxide
DTT	Dithiothretiol
EBD	Estrogen Receptor Ligand Binding Domain
EBD*	Mutant Estrogen Receptor Ligand Binding Domain
EDTA	Ethylendiaminotetraacetic acid

EFS	Event free survival
ES	Embryonic Stem cells
ET	RecE and RecT mediated homologous recombination in *E. coli*
Flp	the site specific recombinase encoded by the 2 micron plasmid of the yeast Saccharomyces Cerevisiae
FRT	the Flp recognition target sequence
GBD	Glucocorticoid Receptor Ligand Binding Domain
GBD*	Mutant Glucocorticoid Receptor Ligand Binding Domain
Hygro	Hygromycin phosphotransferase
HRX	Human Trithorax gene (synonimous of *MLL*)
HTRX	Human Trithorax gene (synonimous of *MLL*)
IRES	Internal Ribosomal Entry Site
kb	Kilobase pairs of DNA
kDa	Kilodalton
LacZ	β-Galactosidase
LBD	Ligand Binding Domain
LBD*	Mutant Ligand Binding Domain
LIF	Leukemia inhibitory factor
loxP	The Cre recognition target DNA sequence
MEFs	Mouse Embryonic Fibroblasts
MLL	MIxed LIneage leukemia gene
neo	Neomycin phosphotransferase
Ori	Origin of replication
pA	polyadenylation site
PAC	P1 based artificial chromosome

PBS	Phosphate buffer saline
Pc	Polycomb gene
PcG	Polycomb group
PGK	Phosphoglycerate kinase
PHD	Plant Homeo Domain
PCR	Polymerase Chain Reaction
RT	Reverse transcription
sA	Splice Acceptor element
SET	Su(var)3-9-Enhancer of Zeste-Trithorax
SSR	Site Specific Recombinase
SSRT	Site Specific Recombinase target site
SV40	Simian Virus t40 Antigen promoter
TAP	Tandem Affinity Purification
TnpI	The site specific recombinase encoded by the transposon Tn4430
TopoII	Topoisomerase II
TRT	The TnpI recognition target DNA sequence
Trx	Trithorax gene
TrxG	Trithorax group
wt	wild type
X-gal	5-Bromo-4-chloro-3-indolyl-β-D-galactopyranoside
YAC	Yeast Artificial Chromosome

INTRODUCTION

I

Chromosomal translocations in human leukemias

I.1 Introductory remarks

Chromosomal aberrations are a hallmark of many neoplasias. They were first connected to cancer at the end of the 19th century, with von Hansemann's thorough investigation of cell division in malignant tumours. Starting in the middle of the last century, karyotyping techniques gradually became available which started to identify recurrent chromosomal abnormalities in a variety of human cancers. This led to the discovery in 1960 of the Philadelphia chromosome (Ph) in chronic myelogenous leukemia (CML), the first karyotype abnormality consistently associated with a disease condition, and since then a paradigm in the study of chromosomal translocations in human leukemias. The next decisive leap forward came at the beginning of the 1970s with the introduction of chromosome banding techniques (Caspersson et al., 1970). These enabled one to identify each individual chromosome unequivocally, and constituted a prerequisite for the precise characterisation of translocations and other aberrations. Soon after it became apparent that the Ph chromosome did not represent a unique case; recurrent chromosomal aberrations were consistently identified in a variety of human neoplasms, mostly leukemias, lymphomas and soft tissue solid tumours.

Four main classes of cytogenetic aberrations can be recognised: interchromosomal translocations, deletions, duplications and inversions. While deletions and duplications result in a loss or gain of genetic material, in both translocations and inversions it is the

rearrangement of a specific locus or loci which carries a pathogenic effect. This class of rearrangements can have two consequences (Look, 1997; Rabbitts, 1994).

A gene, usually a proto-oncogene, can be relocated under the regulatory influence of a new control region, which in turn leads to a derangement in its expression profile. This mechanism is exemplified by translocations involving the c-myc proto-oncogene in Burkitt's lymphoma. This is a B-cell malignancy, which can arise as a consequence of three possible distinct translocations. In all cases, the *C-MYC* proto-oncogene is translocated to either the immunoglobulin (more frequently) or the T-cell receptor gene. Chromosomal fusions happen within the joining or diversity segments of those loci, leading to inappropriate expression of c-myc. A constitutive alteration in the levels of the c-myc protein is believed to shift the overall equilibrium of a network of transcription factors complexes (comprising MAX, MAD and Mxi-1) towards MYC containing, activating complexes. Inappropriate expression of the respective target genes initiates an oncogenic cascade.

A number of other hematological neoplasms are associated with analogous translocations causing disregulation of the rearranged gene. These often encode transcription factors, like *HOX11*, *TAL1* or *RBTN* in acute T-cell leukemias (T-ALLs). These genes are not normally active in T-cells, and their ectopic expression driven by the T-cell receptor gene has been shown to be the key pathogenic event.

A notable example of genes other than transcription factors involved in this kind of rearrangements comes from some of the translocations identified in chronic forms of lymphoma or leukemia. *BCL-2*, which normally protect B and T lymphocytes from apoptosis, is involved in the translocation t(14;18)(q23;q21), where it is juxtaposed to the IgH-joining segment, resulting in its inappropriate expression.

This class of chromosomal translocations, observed exclusively in lymphoid malignancies, has been shown to result almost invariably from mistakes in the V(D)J

recombination process which normally takes place in lymphoid cells to assemble the antigen receptors. This comes mainly from two observations. First, some of the breakpoints which have been molecularly chracterized display canonic heptamer or nonamer sequences which can be recognised by the lymphoid recombination machinery. Second, both translocated chromosomes show an N-nucleotide addition, which is a hallmark of RAG activity. Thus, it seems that mistakes in the normal recombinational activity of lymphoid cells can account for most of these translocations.

Exhaustive analysis of both leukemias and solid tumors has clearly defined that the most common visible outcome of interchromosomal translocations is actually the creation of new chimeric proteins (Rabbitts, 1994). In these cases, the breakpoint occurs in the introns of the genes involved, and the resulting product is a fusion of the domains contributed by each of the two partner genes. Besides the already mentioned Ph chromosome resulting from translocation t(9;22)(q34;q11), the t(4;11)(q21;q23) translocation fusing *MLL* with *AF4* is an example of this type of rearrangement, along with the great majority of all rearrangements affecting the *MLL* gene.

If one considers the whole spectrum of chromosomal aberrations in human cancer, some common themes emerge, raising some of the most outstanding questions which still lie unanswered in our understanding of neoplasia.

First of all, non-random, recurrent chromosomal aberrations have been identified in all human neoplasms which have been analysed in sufficient numbers to allow robust conclusions. Recently, a comprehensive map of all recurrent chromosomal rearrangements was assembled (Mitelman et al., 1997), which analysed 26,523 cases reported in the literature, and identified 215 balanced and 1588 unbalanced recurrent chromosomal aberrations. Balanced aberrations include translocations and some inversions, where the

chromosomal topography is altered without loss of genetic material. Deletions, duplications and inversions accompanied by changes in the amount of genetic material constitute unbalanced aberrations. Balanced aberrations (ie. mostly translocations) exhibit much higher disease specificity, whereas only a few of the unbalanced abnormalities are consistently associated with specific tumors. The reasons for this difference are largely unknown, but it possibly indicates that unbalanced aberrations may not represent the initiating event but rather contribute to later stages of the neoplastic process. The loci affected may encode for a diverse range of proteins, whose disregulation could confer similar selective advantages to diverse tumor types.

The second clear outcome of this comprehensive analysis has been to confirm an older observation, namely that chromosomal translocations are much more frequent in hematologic neoplasms than in solid tumors, among which they occur much more commonly in soft tissue tumors (of mesenchymal origin) than in epithelial ones. This may simply reflect the less advanced state of solid tumor cytogenetics, in which case the introduction of new techniques like multicolour fluoresence *in situ* hybridisation (chromosome painting) should gradually eliminate the difference in translocation frequency between mesenchymal and epithelial neoplasms. However, this discrepancy may well reflect a genuine biological phenomenon, possibly related to the different sensitivity of various tissues to chromosomal rearrangement per se or to its molecular consequences. Needless to say, thorough investigation of this fundamental difference could be extremely rewarding.

The third major conclusion which can be drawn from such a catalogue is that virtually all chromosome bands of the human genome are involved, albeit at different frequencies. Although some regions of the genome are more prone to recombination events and the products of certain rearrangements more often result in selective growth advantages,

clearly a very large number of genes can contribute to the multistage process of carcinogenesis.

The second part of this chapter deals exclusively with interchromosomal translocations, and some of the emerging concepts in their origin and significance.

I.2. Seed versus soil: chimeric fusion proteins and the question of specificity

From a thorough overview of the data gathered so far, it appears that most clinically manifested leukemic translocations involve genes which are required for normal hematopoiesis. This conclusion is largely based on *in vivo* experiments, in which genes translocated in human leukemias have been knocked out in the mouse and found to be essential for various aspects of blood development. It should be emphasised that, although this correlation is now considered obvious and predictable, it need not be necessarily so (Orkin, 2000). Genes disrupted in human leukemias could have been for example just related to the master regulators of hematopoiesis, belonging to the same protein family. Instead, the precision of this correlation clearly points to an intimate relationship between perturbed hematopoiesis and leukemogenesis. As the list of leukemogenic rearrangements continues to expand, an unprecedented opportunity unfolds to understand the physiological basis of blood development and its pathological derangements. For chimeric fusion proteins constitute natural experiments in shuffling together domains of different proteins, often impinging on disparate signaling pathways affecting growth, differentiation and apoptosis.

Animal studies to investigate the normal function of genes disrupted by specific chromosomal translocations have provided evidence as to the specificity of various translocations with respect to the disease phenotypes they are associated with. This central issue framed the "seed versus soil" debate, which usually involves consideration of two alternative scenarios (Barr, 1998; Westervelt and Ley, 1999). According to one hypothesis, the consistent association of a certain translocation with one or a few specific tumours could reflect the fact that this translocation can only happen or exert its pathogenic effect in a

given lineage or at a specific stage of commitment. This could reflect a number of different reasons:

1) The two genes could be accessible for interchromosomal recombination only at certain stages of development, due to the chromatin organisation of their loci or the relative topology of chromosomal territories within the nucleus. Though very intriguing, this is one of the least characterised factors possibly underlying translocation specificity. As an example, the *RET/PTC1* inversion causing thyroid papillary cancer was recently investigated by two colour fluorescence in situ hybridisation coupled with three-dimensional microscopy. Though separated by a linear distance of more than 30 megabases on chromosome 10, this study showed that at least one pair of *RET* and *PTC1* loci was actually juxtaposed much more frequently in thyroid cells than in other tissue types, possibly providing a high-order structural basis for the exclusive occurrence of this rearrangement in thyroid neoplasms (Nikiforova et al., 2000).

2) The recombination machinery which catalyses the translocation reaction could operate only in specific cell types and/or at defined points of development. This is obviously relevant for lymphoid neoplasms (see above), where the activity of the V(D)J recombination apparatus is essential for the translocation to occur.

3) Once the translocation has taken place, the resulting fusion protein needs first of all to be adequately expressed. The availability of the relevant machinery at all levels of the gene expression flux (ie. transcription, post-transcriptional processing, nuclear-cytoplasmic export of the mRNA, translation and post-translational modifications) constitutes an obvious prerequisite.

4) Assuming proper expression, the fusion protein might initiate the oncogenic cascade only under certain conditions. For example, it might need the

simultaneous presence of cofactors (transcription factors, chromatin modifying complexes, signal transducers), whose expression may in turn be tightly regulated both in space and time. Along the same line of logic, potential target genes of the fusion protein might not be always amenable to transactivation or repression, depending for instance on their chromatin accessibility. In such cases, the effect of the fusion protein might be negligible, and its presence tolerated by the cells without further consequences.

5) Finally, the fusion protein might be toxic to certain cell types and/or at specific stages of development. The availability of cellular factors, or extracellular cues, promoting tolerance to these toxic effects would add an additional layer of selection for the permissive lineage and/or cell type.

Thus, in all of the above cases, it would be the lineage ("soil") which actively selects the possible translocation product, by imposing various constraints on its occurrence and/or on its function.

The alternative scenario predicts that it is the translocation which actively determines the lineage. According to this model, a given translocation would randomly occur in a relatively undifferentiated progenitor cell. The resulting fusion protein, by virtue of the intrinsic specificity of its domains, then executes a determined genetic programme leading to expansion of a defined lineage with neoplastic features.

It is clear that these models represent two possible extremes, and that it is reasonable to envision a complex developmental system in which they are not mutually exclusive. There probably are numerous constraints for a specific translocation to occur and to propagate, so that not any translocation can be expected to arise in any tumor. However, it is equally likely that the fusion protein could at least contribute to specifying, if not fully determining, which

developmental decision the target cell will take, among the restricted number of options available at that stage in space and time.

Two approaches have been used to address the issue of specificity, cell culture studies and mouse models.

II

Current approaches to modelling leukemia

II.1 Cell culture studies of fusion proteins

Most studies of cells in culture have underscored the importance of the cellular compartment ("soil") in the transformation process. The strategy of such studies is usually to force expression of a fusion protein in various cell culture systems and assess for the effects on growth and/or differentiation. An obvious limit is that the use of heterologous cell lines, bearing little or no resemblance to the original tumour histotype, can lead to results and interpretations of questionable *in vivo* relevance. However, this very setting can also prove rather useful to dissect the issue of "seed" vs. "soil".

The following examples are representative of the type of conclusions which can be expected from such approaches.

Like many other fusion proteins identified in human leukemias, the BCR-ABL1 product cannot transform NIH3T3 mouse fibroblasts, despite very high levels of expression (Daley et al., 1987). However, transforming activity has been established for certain *BCR-ABL-1* constructs harbouring C-terminal rearrangements, and specific subclones of NIH3T3 cells have been identified which are readily transformed by the BCR-ABL-1 fusion protein, strongly arguing for a complex interplay between distinct domains of the fusion protein and the specificity of the cellular milieu (Renshaw et al., 1995; Shore et al., 1994).

Again from the NIH3T3 system, it emerged that although the fusion protein may itself exert a transactivation function on gene expression, this readout can be of no relevance unless it involves genes of functional importance for a specific lineage.

Twelve genes were identified as being induced in NIH3T3 cells upon expression of the oncogenic fusion product E2A-PBX1, usually associated with pre-B acute lymphoblastic leukemias. Most of these genes are not expressed either in human pre-B leukemia lines, or in *E2A-PBX1* immortalised mouse myeloblasts, both arguably much closer to the original leukemic cells than NIH3T3 fibroblasts (Fu and Kamps, 1997).

The results obtained with the *PML-RARα* translocation are even more telling. As with many other fusion products, it proved impossible to stably express the PML-RARα protein in a variety of cell lines, including the majority of hematopoietic cell lines (Ferrucci et al., 1997). In contrast, high levels of expression were readily achieved in hematopoietic cell lines derived from myeloid leukemias. The introduction of inducible expression systems revealed that this fusion protein induces apoptosis in the majority of cell lines tested. The myeloid lineage seems to be exquisitely resistant to this toxic effect, providing a rational framework for the specific association of this translocation with acute promyelocytic leukemia (APL).

The use of more relevant cell culture systems can certainly obviate some of the problems described, as for example in the studies which investigated the potential of the PML-RARα and FUS-DDIT3 fusion proteins in myeloid and adipocytic cell lines, respectively (Grignani et al., 1993); (Kuroda et al., 1997). Nonetheless, cell lines pose a conceptual problem for a meaningful interpretation of this kind of experiments. In fact, being already transformed or else immortalised, they provide a dubious background against which to test the genuine oncogenic potential of specific fusion proteins.

II.2 The rationale behind mouse models

Although cell culture experiments can provide interesting clues for a preliminary understanding of the leukemogenic process, especially at the biochemical level, it is clear

that a thorough characterisation of cancer can only come from suitable animal models. There are several reasons for this:

1) Though initially triggered in a single cell, cancer invariably becomes a disease of the organism, in which a myriad of host-tumor interactions contribute to the final biological outcome. Arguably, this is the biggest challenge for phenotypic characterisation, and yet it is essential for a complete and genuine understanding of the disease.
2) Cancer is a dynamic process, where several genetic and epigenetic lesions appear and are selected over time. This temporal and evolutionary aspect can be only partially recapitulated in cell culture systems.
3) The ability to metastasise (and in the case of leukemias to infiltrate different tissues) is the single most relevant feature of a tumour, certainly from the standpoint of clinical management. Although this process can greatly benefit from pilot in vitro experiments, it can only be fully addressed in the context of the whole organism.
4) The generation of meaningful mouse models of human cancers constitutes a platform for the identification of modifier loci by genetic screening methods. The characterisation of cancer risk modifying genes will help elucidate the complex molecular network which underlies cancer development. It will also constitute a key step for the coming of age of preventive and predictive oncology.
5) Finally, the availability of mouse models which faithfully mirror human neoplasms is a prerequisite to test the feasibility and efficacy of novel therapeutic approaches.

The next sections summarise the various methodological approaches which have been used to engineer mouse models of human leukemias, highlighting their respective strengths and

weaknesses in the light of the results obtained and the paradigms which are beginning to emerge.

II.3 Standard transgenic approaches

Standard transgenesis approaches were the first ones to be applied. In this type of experiment, a cDNA coding for the desired fusion protein under the control of a suitable promoter element is integrated into the mouse genome by injection into fertilised oocytes. As with any transgenic experiment, the overall expression levels achieved with a given promoter are strongly influenced by the site of genome integration and sometimes by the number of copies of the construct which have been integrated. The site of integration is a particularly relevant variable, as local control regions can override the transgene's own promoter elements causing either inappropriate silencing or inappropriate expression. Hence the need to characterise multiple founder lines to identify the most faithful pattern of expression.

An additional concern comes from the possibility of insertional mutagenesis following integration of the transgene, resulting in disruption of an unrelated gene or alteration of its chromatin environment. In either case, a spurious source of phenotypic variation can confound the interpretation of results.

While these limits are inherent in any transgenic experiment, in the case of translocation models, this approach has additional shortcomings.

First of all, in the vast majority of cases, interchromosomal translocations constitute acquired somatic mutations which occur sporadically. In the first transgenic models of leukemias, on the contrary, the fusion protein was expressed throughout the development of the organism. In some cases, this resulted in embryonic lethality, demonstrating a toxic role of the fusion product for certain lineages and certain phases of development. This occurred

for example in experiments in which the expression of the BCR-ABL1 fusion protein was driven by the BCR promoter itself (which would have seemed at first to be a relatively faithful approach to model CML) (Heisterkamp et al., 1991).

Even in those cases in which the animal survived, and a leukemia eventually ensued, a conceptual flaw remained, in that all the cells in the whole lineage from which the disease eventually originated were constantly confronted with the presence of the translocation product; an artificial situation whose outcomes might not be easily applicable to the human disease.

In order to partially overcome these problems, more sophisticated transgenic approaches were pursued in which the regulatory element was carefully selected on the basis of the suitability of its spatio-temporal expression profile to the lineage and/or developmental stage thought to be important for leukemogenesis. This type of design can avoid general expression of the fusion protein throughout the organism. However, it still results in the activity of the translocation protein in all cells in which the regulatory element is switched on. Again, this exposes the animal to a "mutational burden" which has no counterpart in human pathogenesis.

A second important limitation is that in transgenic models of leukemias, only one of the two fusion proteins is expressed. For some neoplasms, the relevance of both translocation products to the development of the disease has been questioned, mainly because only one of the two derivative chromosomes is consistently detected. Such findings must be interpreted carefully, however, as they might simply indicate that one of the two fusion proteins contributes to the initial transforming steps, but becomes then dispensable once the cell has acquired additional genetic lesions, which is the stage at which most leukemias come to clinical attention. And even if only one fusion protein were instrumental in the development of the disease, the reciprocal product could play a role in the leukemic

phenotype, as demonstrated in at least one mouse model, where the presence of both translocation products affected both the penetrance and the phenotype of the disease, probably via a heightened predisposition to accumulate additional genetic lesions (Pollock et al., 1999; Zimonjic et al., 2000).

Third, transgenic models do not recapitulate the relative gene dosage of a cell which has undergone an interchromosomal translocation. Such a cell will express only one wild type allele for each of the two genes involved, and this could well contribute to the transformation process if the function of these genes is dosage sensitive. In transgenic models, in contrast, the fusion product is expressed in addition to the normal alleles of both target genes.

Fourth, there has been very little investigation into the long range consequences that interchromosomal translocations can have on genes other than the ones which are actually interrupted. In recent years, functional meaning has been attributed to the non random positioning of chromosomal territories in interphase nuclei. Together with observations which have identified distinct subsets of subnuclear compartments or bodies, this has led to a fluid model of nuclear organisation in which specific domains are defined by the local concentration of specific proteins (chromatin regulators, transcription factors, DNA and RNA processing enzymes, nuclear signal transducers). It is not hard to imagine how the violent perturbation resulting from the joining of two unrelated chromosomal territories could have pleiotropic effects on nuclear physiology. Furthermore, it is not even necessary to invoke such higher order interactions; a variety of long range enhancer-promoter circuits are likely to be affected by interchromosomal translocation. The phenotype we eventually observe could indeed be the combined result of a major oncogenic hit (the fusion protein) and a concomitant array of more subtle molecular changes.

Despite these drawbacks, some previous transgenic attempts did result in useful models of leukemia. Transgenic models of the *PML-RARα* translocation are here summarised as a prototypic case study, since the differences between the various models produced constitute a useful framework from which to explore in vivo the question of "seed vs. soil".

II.3.1 The *PML-RARα* paradigm

Four different transgenic models have attempted to recapitulate acute promyelocytic leukemia (APL) in the mouse. At one end of this group stand the two experiments which yielded leukemia in the mouse, albeit with different phenotypes. At the other extreme lies the interesting case in which no leukemia was observed, bur rather an impairment in myelopoiesis. Adding one more colour to the puzzle, in a fourth model, the expression of the fusion protein resulted in neoplasm of a different tissue (hepatic neoplasms).

In the two successful models, the same PML-RARα fusion protein was expressed under the control of either cathepsin-G or MRP8 promoter (Brown et al., 1997; Grisolano et al., 1997). The cathepsin-G construct restricts expression of the cDNA only to the promyelocytic compartment, while the MRP8 gene is expressed throughout myeloid development. In both cases, a myeloid leukemia recapitulating many aspects of the human disease did develop. However, the frequency of occurrence and the preleukemic phenotype were dramatically different. All of the mice expressing PML-RARα under the control of cathepsin-G showed myeloid expansion with terminal differentiation, and 30% of those animals developed acute leukemia; interestingly, the leukemic blasts could not be differentiated in vitro with all-trans retinoic acid (ATRA), one of the main features of APL cells in humans.

Mice expressing the same fusion protein under the control of the MRP8 gene showed a substantially normal bone marrow with minor myeloid abnormalities. In 5% of the animals a promyelocityc leukemia developed with a complete block in granulocyte maturation (in contrast to the cathepsin-G animals where full myeloid maturation was observed concomitant to the leukemia). Strikingly, these blasts could be differentiated in vitro by ATRA treatment.

The difference in differentiation phenotype between these two models (partial versus complete differentiation block in cathepsin-G versus MRP8 mice, respectively) could be due to the fact that while cathepsin-G expression is restricted to the promyelocytic compartment, the MRP8 gene stays active throughout terminal differentiation. An even stronger support for the importance of the targeted "soil" in determining the leukemic phenotype comes from the analysis of the "unsuccessful" model. In this case, the PML-RARα protein was placed under the control of the CD11B promoter (Early et al., 1996). The CD11 gene is expressed in granulocytes, but not in myeloid precursors. No leukemia was observed, but rather a mild impairment of myelopoiesis, suggesting that this fusion protein can only unfold its oncogenic potential at specific stages of myeloid development.

The parallel importance of the "seed" comes from experiments in which the *PLZF-RARα* fusion (the product of one of the variant translocations leading to APL) was expressed under the same cathepsin-G promoter described above (He et al., 1998). While the preleukemic phenotype was similar to the CG-*PML-RARα*, penetrance was in this case 100%, and in addition the differentiation block appeared less severe. While differences in transgene integration site and/or dose could theoretically explain these findings, it is intriguing to speculate that this variation is specifically due to the different fusion proteins employed.

Finally, mice in which the promyelocytic leukemia fusion protein was expressed under the control of the metallthionein-1 promoter developed hepatic preneoplastic and neoplastic lesions, showing that more than one lineage appears sensitive to the transforming effect of PML-RARα (David et al., 1997).

II.4 Knock-in approaches

A more sophisticated variation in leukemia modelling involves targeting of the fusion protein cDNA to an endogenous locus by homologous recombination rather than random integration. As for standard transgenic approaches, the endogenous locus is selected for the suitability of its spatio-temporal expression profile to the lineage and/or developmental stage thought to be important for leukemogenesis. These type of designs avoid most of the general problems encountered with standard transgenesis, namely variegation of expression and multiple copy integration. All other drawbacks are still relevant though, with an added complication resulting from the necessary disruption of the endogenous targeted allele, which obviously severely restricts the choice of possible loci.

A step forward in leukemia models was achieved in 1996, when Rabbitts and coworkers applied a more sophisticated "knock-in" strategy, in which the 3' terminal portion of the cDNA of one of the two target genes was inserted into the endogenous locus of its translocation partner at the site of the breakpoint, yielding expression of the fusion protein under the control of the endogenous upstream partner. Four studies have been carried out with this approach. Rabbitts et al. first applied it by knocking the cDNA of the Af9 gene into the *Mll* locus (Corral et al., 1996). Analogous experiments were then performed with BCR-ABL1, CBFB-MYH11, and AML1-ETO (Castellanos et al., 1997; Castilla et al., 1996; Yergeau et al., 1997; Okuda et al., 1998).

An extensive account of the *Mll*-Af9 leukemia model is given in chapter IV.8.1. Here, the findings of these four studies are collectively analysed in the context of their modelling potential.

There are several advantages to this kind of approach. First, it avoids the intrinsic caveats of standard transgenesis. Only a single copy of the fusion protein is expressed, and its expression is predictably controlled by the endogenous locus of the 5' partner gene. This at least partially mirrors the human disease. More, the very nature of the technique employed to generate the mice (blastocyst injection of modified ES cells) enables the analysis of chimeric mice, with varying degrees of contribution from the mutated ES cells. Only a subset of the lineages in which the endogenous locus is active will therefore express the fusion protein, leading potentially to a lower oncogenic burden and a more faithful model. On the contrary, in transgenic experiments, chimeras are only rarely generated (10-30% of cases), and more importantly the percentage of chimerism cannot be controlled.

However, this methodology also has its shortcomings. Again, only one of the fusion proteins is expressed, and the overall gene dosage, though closer to the human disease, is still not fully recapitulated. Investigation of the effects of disrupting long-range chromatin interactions also falls beyond the possibilities of such models.

Another very important limitation actually coincides with one of its potential merits, namely the use of the endogenous promoter of the 5' gene to drive expression of the fusion protein. This can be a very powerful approach for genes with specific expression restricted to the lineage from which the leukemia originates. However, many of the genes translocated in human leukemias are often normally expressed in multiple cell types, sometimes from early on in development. A more global expression of the fusion protein can have diverse effects, as exemplified by the case of *Mll*-Af9 on the one hand, and Aml1-Eto and Cbfb-Myh11 on the other.

Although the *Mll* promoter is active in a wide spectrum of tissues from early development, only leukemias developed (Corral et al., 1996), providing an adequate system with which unravel the "seed vs. soil" issue. On the contrary, in both Aml1-Eto and Cbfb-Myh11 experiments, modified ES cells were excluded from the hematopoietic lineage of chimeric animals and no leukemia developed. Upon germline transmission, mice died at midgestation because of grossly impaired hematopoiesis. Thus the widespread and early activity of the endogenous promoters had a catastrophic effect in these two models. The similarity of these phenotypes with the ones of Aml1 or Cbfb homozygous mutants strongly suggested that the fusion proteins exert a dominant negative function to disrupt hematopoiesis. Furthermore, the fact that with these approaches the expression of the fusion protein is dependent on the regulatory pattern of the endogenous promoter precludes the possible use of temporal regulation to dissect the role of the translocation product at different stages throughout development and hematopoiesis.

To conclude, some previous approaches have resulted in very relevant mouse models which partially mirror some aspects of the leukemogenic process. None of them however was designed to be able to exactly recapitulate the human disease. Cre-loxP technology is ideally suited for this purpose, and will be described in the next section.

II.5 Cre-loxP technology based approaches

II.5.1. Cre recombinase in genome engineering

Cre recombinase (cyclization recombination) is a 38KDa site specific recombinase (SSR), isolated from bacteriophage P1, which catalyses recombination between two 34 bp recognition elements, called loxP sites. The loxP site is composed of two 13 bp inverted repeats flanking an asymmetrical 8 bp spacer sequence, which confers directionality to the recombination reaction. Thus, depending on the relative orientation of the two loxP sites,

Cre can catalyze either deletion or inversion of the intervening sequence. If the two loxP sites are located on different chromosomes, the product is an interchromosomal translocation.

The use of this system for developing mouse models of human leukemias that are associated with interchromosomal translocations relies on the introduction in the mouse germline, through ES homologous recombination, of loxP sites in the particular introns of the two genes involved in the human translocation. Upon Cre expression, recombination between the loxP sites is expected to result in the desired chromosomal translocation. The main advantages of the system are the following:

1) As in the human disease, both derivative chromosomes are created in the target cell, potentially resulting in the expression of both fusion proteins and therefore in a more faithful phenotype.

2) The dosage of the two genes perfectly mimics the situation of a leukemic cell. This could be particularly important in the case of the *MLL* gene, whose haploinsufficiency in hematopoiesis has been postulated to play a role in initiating leukemogenesis.

3) Any potential effect due to the aberrant joining of distinct chromosomal domains and/or to the concomitant disruption of long-range chromatin interactions is also recapitulated.

4) The translocation is present in only a subset of somatic cells. Therefore, it cannot interfere with normal embryonic development and adult hematopoiesis, and enables assessment of the functional relevance of the translocation in the presence of ongoing, normal hematopoiesis.

5) By selecting appropriate Cre expressing mouse lines, it should in principle be possible to define precisely in which lineages and at what stages of blood

development the translocation exerts an oncogenic effect. Plus, by tightly regulating the expression and/or activity of Cre recombinase in a certain lineage, it could be possible to answer one of the fundamental and most open questions in leukemia (and cancer) biology, namely, how often and/or in how many cells a potentially oncogenic rearrangement must occur in order to give rise to an overt malignancy.

Potential disadvantages of this approach mainly include the efficiency at which Cre catalyses the interchromosomal translocation *in vivo* (see below), and the necessity that in the mouse the two genes are in the same centromere-telomere orientation as in humans. Otherwise, Cre recombination would result in a dicentric and an acentric chromosome, leading to the death of the targeted cell.

Recently, interchromosomal translocations were achieved *in vivo* in the mouse. The seminal relevance of these results can be fully appreciated in the context of the substantial research activity which has over the years explored the potential of SSRs (in particular Cre) as tools in chromosome engineering. FLP-mediated translocation between homologous chromosomes in Drosophila was one of the first applications of site specific recombination in genome engineering (Golic, 1991). However, for a long time it remained an open question whether an SSR driven non-homologous translocation could be achieved in the mouse at a detectable frequency. It was clear from other ("easier") genome engineering exercises that efficiency dropped as a function of the distance between intrachromosomal SSR target sites (Ramirez-Solis et al., 1995; Ringrose et al., 1999; Zheng et al., 2000). Additionally, evidence indicating that reasonable efficiencies of SSR translocations require the pairing of homologous chromosomes achieved during mitosis (Golic and Golic, 1996), or meiosis (Herault et al., 1998) was obtained in both flies and mice. A further source of doubt arose with the emergence of the chromosomal territory model of nuclear organisation (Cremer and Cremer, 2001; Lamond and Earnshaw, 1998; Zink et al., 1998), which led one to wonder

how the two SSR target sites, buried in distinct and possibly distant chromosomal territories, would ever manage to be brought together for Cre recombination to take place in interphase. Hints for optimism, but also pessimism, came from successful pioneering experiments with Cre mediated translocations between non-homologous chromosomes in mouse ES cells (Smith et al., 1995; Van Deursen et al., 1995), as these studies provided a first indication of the frequency of this event. Smith and coworkers used an approach in which the translocation event was positively selected through the reconstitution of the Hprt minigene in a hypoxanthine phosphoribosyltransferase (HPRT) deficient ES cell line. The translocation efficiency was on the average 5×10^{-8}. However, Cre expression was achieved with a transient transfection strategy, hence the above figure is likely an underestimation of the effective translocation frequency among cells expressing Cre. This was confirmed by the other study exploring interchromosomal translocation in ES cells (Van Deursen et al., 1995). The authors obtained Cre mediated translocations without exerting any selective pressure, and the translocation was detected by nested PCR. On the basis of a PCR serial dilution analysis, and taking into account the transfection efficiency of the Cre expressing plasmid, the translocation was estimated to occur in 1 in 1200-2400 ES cells expressing the recombinase. Still, despite the considerable achievement represented by these two works, the fact that the experiments were limited to cells in culture gave rise to legitimate doubts as to whether the result could be recapitulated in a living mouse.

The breakthrough occurred when two studies reported successful Cre mediated interchromosomal translocation *in vivo* in an attempt to model the *MLL*-AF9 and the AML1-ETO leukemias, respectively (Collins et al., 2000; Buchholz et al., 2000). In both cases, the translocations were detected by PCR. Failure to detect them by Southern hybridisation indicates that as expected translocation is a rare event, and in one study it was estimated by

semiquantitative PCR to occur in 1 in 10000 to 1 in 1000000 cells (Buchholz et al., 2000). Will these frequencies be sufficient to induce leukemogenesis? It is currently unclear, as none of the two studies has yet reported the occurrence of leukemia, a necessary result to conclude that this approach is technically superior to all others generated so far. When assessing the overall probability that this approach could result in a useful model of tumorigenesis, the following aspects need to be taken into account.

First, one decisive factor is the timing and level of Cre expression. In this regard, the mouse offers an advantage over cell culture approaches, in that by appropriate breeding the two translocation prone chromosomes and Cre expression can be combined in every cell of the animal. While this would be by no means a homogeneous population of cells (for example the chromosome territory organization could be substantially different among tissues), it does constitute a good starting point to assess translocation frequency. Furthermore, the fact that in one of the two studies described above (Collins et al., 2000) the translocation was not observed in the bone marrow, probably due to insufficient Cre expression in this compartment, is a clear demonstration that appropriate Cre mouse lines are needed before drawing any conclusion on the feasibility of of Cre-loxP based approaches..

Second, Cre recombination is a reversible reaction, and it is therefore possible that the reverse translocation event, which would restore the two normal chromosomes, would happen with a similar frequency as the direct rearrangement. Although the paradigm holds that the cell hit by the translocation acquires a selective growth advantage through the action of the fusion protein, many factors would influence the final outcome in the face of ongoing Cre activity: the rate of the reverse translocation; the half life of the fusion protein; the doubling time of the hit cell and its variation as a function of the translocation; and the timeframe needed by the fusion protein to derange cellular homeostasis. Most of these

aspects are completely unknown or poorly chracterised. Thus, the best way to address the problem of reversibility is to use conditional forms of Cre, which would deliver only transient pulses of Cre activity and possibly avoid the reverse reaction. Important variables to consider, when using this approach, are the half life of Cre (which in turn could vary for different cell types) and the pharmakokinetics of the molecule administered to switch on the enzyme (usually a nuclear hormone receptor ligand analogue). This last parameter will affect the overall size of the target cell population, and will thus be a central component in the trade-off between a constitutive Cre approach (with the drawback of reversibility) and the conditional Cre approach (with the drawback of a smaller cohort of cells able to translocate).

Third, one central unresolved issue of cancer biology is the frequency with which a certain rearrangement and/or mutation must occur within an organism in order to give rise to a tumor, a factor which is likely to differ considerably according to the kind of rearrangement and the cell type involved. This is in the end the crucial parameter which will determine the success of Cre-loxP based approaches given the frequencies reported above. Recently, a cancer mouse model was reported using, for the first time *in vivo*, a "hit and run" approach to mutagenise the *Ras* oncogene (Johnson et al., 2001). Aiming at recapitulating the sporadic occurrence of carcinomas in humans, the authors engineered mice harbouring a mutant copy of the *Ras* oncogene in a silent configuration ("hit" step), capable of generating an active mutant allele following homologous recombination ("run" step) either in an intrachromosomal reaction or in an unequal sister-chromatid exchange. The novelty lies in the fact that the "run" step was allowed to occur *in vivo* relying on the spontaneous rate of recombination between duplicated genomic sequences, which has been estimated to range between 10^{-3} and 10^{-7} per cell division (Hasty et al., 1991; Seperack et al., 1988). All mice developed tumours with a wide range of phenotypes. Although these results may not be directly related to the Cre-loxP translocation models discussed above since the oncogenic

products are very different, they constitute initial evidence that relatively low mutagenesis frequencies can result in tumorigenesis.

II.5.2 Conditional Cre expression and activation: current strategies

Throughout the remarkable story of Cre applications to mouse genome engineering, one of the most significant advances has been the development of regulatable forms of Cre obtained by fusing it to ligand binding domains (LBDs) of nuclear hormone receptors. Pioneering experiments with Flp and Cre (Feil et al., 1996; Kellendonk et al., 1996; Logie and Stewart, 1995; Zhang et al., 1996) demonstrated that the ligand-dependent properties of steroid hormone receptors could be imposed upon recombinases, allowing regulation of their enzymatic activities by administration of the appropriate ligand. In the absence of ligand, the LBD, through its interaction with the ubiquitous heat shock protein 90 (HSP90), traps Cre in an inactive complex. Upon ligand binding, the Cre-LBD fusion is released from this complex and can now bind the loxP sites, where it catalyses the recombination reaction.
This approach opened the way to tightly controlled spatio-temporal somatic mutagenesis in the mouse (Schwenk et al., 1998). The spatial aspect of regulation is achieved by restricting expression of Cre (or Flp) to selected tissues by placing the recombinase expression under the control of suitable promoters. This strategy can also partially accomplish a temporal control, if the promoter chosen is active in the desired tissue only at a specific stage of development. However, true temporal regulation can only be achieved with the Cre-LBD fusions described above. To be useful for *in vivo* mouse mutagenesis, Cre-LBD fusions must be activated only by exogenous ligands, as responsiveness to physiologically present steroids would undermine the core of the whole strategy. To this end, mutant forms of LBDs (LBDs*), which are sensitive only to exogenous steroid analogues were developed. In cell

culture assays, it has been possible to predictably affect the efficiency of recombination by varying the doses of ligands administered.

The steroid receptor LBDs which have so far been employed are those from the estrogen, the progesterone and the glucocorticoid receptors (ER, PR and GR). For the ER LBD (EBD), several mutations have been described. The G521R mutation in the human EBD (which substitutes a glycine with an arginine, also called Cre-ER^{T}) resulted in a more than 10000 fold reduction of the affinity for the endogenous ligand β-erstradiol (Danielian et al., 1993; Feil et al., 1996). The mutation also reduced the affinity for the synthetic antagonist 4-OH-tamoxifen by about 100 fold (Schwenk et al., 1998), resulting in the need to treat mice with high doses of tamoxifen. While this can be irrelevant in many experimental setups, it clearly limits the usefulness of this mutant for more specialised applications, such as for example induction of recombination *in utero* (Danielian et al., 1998).

Another Cre-EBD fusion was recently characterised and shown to be more sensitive to 4-OH-Tamoxifen than the previous one, respectively four-fold in cultured cells and ten-fold in the mouse epidermis (Feil et al., 1997; Indra et al., 1999) . This mutation has been called Cre-ER^{T2} and contains three mutations in the human EBD, glycine to valine at position 400, methionine to alanine at postion 543 and leucine to alanine at position 544. Importantly, as in the case of Cre-ER^{T}, no background recombinase activity was observed.

As for the GR LBD (GBD), a mutation was identified (isoleucine to threonine at position 747) which results in complete unresponsiveness to endogenous ligands like cortisol or corticosterone, while synthetic analogues like dexamethasone are still able to induce transactivation (Brocard et al., 1998; Roux et al., 1996). As expected, higher dexamethasone concentrations (100 fold more) were needed for this effect, if compared to the natural GBD.

The importance of tightly regulating Cre activity in the mouse was recently underscored by a report of Cre-mediated illegitimate chromosome rearrangements in transgenic mouse spermatids (Schmidt et al., 2000). In this study, Cre was expressed in postmeiotic spermatids under the control of the protamine promoter. All transgenic males were sterile, while males carrying equivalent transgene levels harbouring an inactive form of Cre were normally fertile. Analysis of embryos from sterile Cre-transgenic mice demonstrated the occurrence of chromosome rearrangements which led to abortion with 100% penetrance. Thus, directly expressing Cre in spermatids apparently poses an insurmountable recombinational burden, and it is hypothesised that inappropriate recombination could depend on the presence of spurious loxP sites in the mouse genome. These findings are particularly relevant for all mouse models using Cre to induce various chromosomal aberrations, as the concomitant presence of undesired rearrangements could confound the phenotypic analysis.

In conclusion, conceptual as well as practical reasons strongly advise the generation of inducible-Cre mouse lines as robust and reliable tools for contemporary mouse genetics.

III

Translocations involving the *MLL* gene

III.1 Promiscuity and common themes

Translocations involving band 11q23 are found in about 5% of patients with acute myeloid leukemia (AML) and 10% of the cases of acute lymphoblastic leukemia (ALL); in addition, they are also present in a large portion of leukemias which express markers of both lineages (mixed lineage or biphenotypic leukemias) (Rowley, 1998). These features are very unusual, since most other translocations are associated with only one leukemic phenotype, though this may be broader than the corresponding developmental stage of normal hematopoiesis.

Another unique feature of band 11q23 translocations is the astounding number of partner chromosomes involved. More than 40 chromosomal loci have been shown to be translocated to band 11q23 in a large variety of hematological neoplasms (mostly, but not exclusively, leukemias). The full weight of this observation became clear when positional cloning of the breakpoint cluster region identified the gene involved, which was termed *MLL* for mixed-lineage leukemia (Djabali et al., 1992; Gu et al., 1992b; Tkachuk et al., 1992; Ziemin-van der Poel et al., 1991). Other names include HRX or HTRX for human trithorax, since this gene was the first discovered mammalian homologue of the *Drosophila* trithorax gene (see below). With the concomitant cloning of many of the translocation partners (to date 18 fusion partners have been characterised), one of the major tasks has been to classify these various translocations in terms of their overall frequency, and on the basis of the distinct phenotypes with which they are associated. Although different clinical statistics show some variability in the relative frequencies, comparison of data from large patients cohorts has established at least two findings: some translocations are by far more common

than others, and specific translocations are associated, in many but not all cases, with a distinct leukemic phenotype. Combined with the important observation that the breakpoint cluster region (BCR) of *MLL* is very short [8.3 kb spanning 6.introns and 7 exons, according to the map and nomenclature described in (Nilson et al., 1996)], meaning that much the same aminoterminal portion of *MLL* is retained in all fusion proteins, this has led to the current paradigm holding that the domains contributed by the different translocation partners play a pivotal role in lineage determination. This has been partially confirmed by both cell culture and animal studies (see below). However, contradictory evidence from some studies, together with the observation that the correlation between a specific rearrangement and a certain leukemic phenotype is far from complete, argues that the molecular explanation is likely to be more complex. The following overview of the most frequent rearrangements with the corresponding phenotypes supports this conclusion. The main features of the *MLL* translocation partners most relevant for this discussion are summarised in table 1, largely drawn upon the recent conslusions of the European Union Concerted Action Workshop which provided a comprehensive analysis, unprecedented for its size and significance, of 550 cases of leukemias and myelodysplastic syndromes (MDS) associated with *MLL* rearrangements (Johansson et al., 1998; Secker-Walker, 1998).

While on the whole *MLL* rearrangements account for 5-10% of all acute leukemias and myelodysplastic syndromes, they tend to occur with particular frequency in two clinical settings, infant leukemias (below 1 year of age) and secondary leukemias arising in patients treated with DNA topoisomerase II inhibitors for previous neoplasms. The infant leukemias will be discussed in the section dedicated to the t(4;11)(q21;q23) translocation. As for the iatrogenic leukemias, about 5 to 10% of cases of any given *MLL* rearrangement are associated with prior treatment with topoisomerase II inhibitors.

Table 1: Overview of MLL translocation partners

Translocation partner	Locus	Frequency	Leukemic phenotype	Domains retained in MLL fusions	Functional properties	Additional remarks	References
AF4 (FEL)	4q21	40%	Acute lymphoblastic leukemia (ALL) of pre-B phenotype with coexpression of myeloid antigens	Transactivation domain; nuclear localization signal; C-terminal domain homologous to *Drosophila* lilli, involved in embryo segmentation and cell size regulation	Knock-out mice have a reduction in the CD4+/CD8+ thymocyte compartment	Founding member of a new family of genes which include LAF-4 and FMR-2.	Gu et al., 1992b Morrissey et al., 1993; Prasad et al., 1995; Isnard et al., 2000; Tang et al., 2001; Wittwer et al., 2001
AF5q31	5q31.1	1 case reported	ALL of pre-B phenotype	Transactivation domain; nuclear localization signal; C-terminal domain homologous to *Drosophila* lilli,	Homologous to AF4 in three regions: the N-terminal domain (62% homology); transactivation domain (57%) and C-terminal domain (48%).		Taki et al., 1999
AF9	9p21-22	27%	Acute myeloid leukemia (AML)	C-terminal domain homologous (82%) to the transactivation domain of AF19; nuclear localization signal.	Mll-Af9 knock-in mice have expansion of the myeloid lineage and develop AML	Homologous to yeast protein TFG3, member of SWI/SNF , TFIIF and TFIID complexes.	Nakamura et al., 1993; Cairns et al., 1996; Corral et al., 1996; Dobson et al., 1999.
AF19 (ENL)	19p13.3	< 12%	AML and ALL	C-terminal transactivation domain; nuclear localization signal	Mice infected with a retrovirus coding for the Mll-Af19 cDNA develop AML. When fused to Mll, the 84 C-terminal residues of Af19 are necessary and sufficient for immortalization.	Homologous to yeast protein TFG3, member of SWI/SNF , TFIIF and TFIID complexes.	Tkachuk et al., 1992; Rubnitz et al., 1994; Lavau et al., 1997; Slany et al., 1998; Cairns et al., 1996
AF10	10p12	4-6%	mostly AML	leucine zipper;	The extended PHD finger (ePHD), excluded from MLL fusions, mediates homo-oligo merization *in vitro*.	Homologous to AF17. It has a centromere-telomere orientation opposite to MLL, hence the observed rearrangements involve at least an inversion and a translocation.	Chaplin et al., 1995a; Chaplin et al., 1995b; Chaplin et al., 2001; Linder et al. 2000
AF17	17q21	1-1.5%	Mostly AML	leucine zipper;		Homologous to AF10.	Prasad et al., 1994
CREB binding protein (CBP)	16p13	overall rare	AML and myelodisplastic syndrome (MDS) secondary to Topoisomerase II inhibitors	the whole of CBP except the nuclear hormone receptor interacting domain (NID): CREB binding domain, PHD finger, Bromo domain	Transcriptional coactivator; it possesses histone acetyl transferase activity.	It can also be translocated to MOZ gene in t(8;16) acute myeloid leukemias.	Taki et al., 1997;
p300	22q13	1 case described	AML	Breakpoint is further downstream than with CBP, hence only are included in the fusion.	Transcriptional coactivator; it possesses histone acetyl transferase activity.	Highly homologous to CBP, however they are not functionally equivalent.	Ida et al., 1997

These two clinical settings both strongly support the notion that *MLL* translocations act with extremely short latency. For example, the latency between treatment with topoII inhibitors and leukemia development is almost invariably in the range of few months to a few years, in contrast with secondary leukemias associated with other aberrations (most commonly 7q- or 5q-), which follow treatment with alkylating agents or radiation and for which the latency is usually 5 to 10 years, and often longer.

At present, it remains to be proven that *MLL* rearrangements occurring outside of these two selected patient cohorts (infants and treatment related) also cause leukemia with the same short latency. If this were indeed the case, it could indicate that derangement of *MLL* function is a particularly powerful oncogenic event, which therefore does not need the accumulation of secondary transforming hits. Conversely, if short latency were a specific feature only of these two settings, the effect could possibly be attributed to the *AF4* contribution (for the infant leukemias, where *MLL-AF4* translocations are exceedingly common), or to concomitant treatment-induced genetic lesions, in the case of iatrogenic secondary leukemias.

The identification of so many partner genes poses a puzzle in terms of the mechanism of leukemogenesis. At first sight, there are no obvious properties shared by all partners, and therefore a unified model for MLL leukemogenesis is hard to envision. However, some common themes can be outlined (Dimartino and Cleary, 1999).

First, some translocation partners do share important similarities. This is the case for ENL and AF9, which are 82% identical at their amino- and carboxytermini, and are both homologous to a yeast protein, TFG3 (also called TAF30 or ANC1), a member of both the SWI/SNF chromatin remodelling complex and the TFIIF and TFIID transcription complexes

(Cairns et al., 1996).This in turn has lead to the hypothesis that *ENL* and *AF9* might contribute to leukemogenesis by recruitment of the human equivalents of the SWI/SNF class of chromatin remodelling activities. A transactivation domain has been characterised for ENL in cell culture experiments (see below) and is suspected also for AF9 .

Along the same line of similarity among partners, *AF4* and *AF5*q21 are actually members of the same gene family. Also for AF4, in analogy to ENL, a transactivation domain has been postulated on the basis of GAL4 fusion assays, thus constituting another possible element of commonality in the leukemogenic process.

Similarly, both *CBP* and its highly related cognate gene p300 have been shown to be translocated to *MLL*, although *CBP* is involved much more often. CBP is a central protein in chromatin regulation. It possesses histone acetyl transferase activity (HAT) and has been shown to interact with a variety of transcription regulatory proteins through its multidomain modular structure. In particular, it interacts with nuclear hormone receptors through an aminoterminal domain (NID), with sequence specific transcription factors through two cysteine/histidine rich regions, one of which was originally identified as the CREB binding domain, and the other as the interaction domain with the E1A protein, and with transcriptional coactivators (like SRC-1) via its carboxyterminal domain. Plus, it also features centrally located bromo and PHD domains. The relevance of the CBP/p300 pair of proteins to cancer development is demonstrated by several findings. Missense mutations and whole gene deletions have been found in many solid tumors. Loss of one *CBP* allele results in Rubenstein-Taybi syndrome, a complex developmental disorder which includes cancer predisposition. And finally, *CBP* can be translocated not only to *MLL*, but also to the MOZ gene (the human homologue of the yeast gene SAS) in the t(8;16) acute myelod leukemias. In *MLL-CBP* translocations, almost the whole *CBP*, except for the NID, is retained in the

fusion protein, suggesting that abnormal recruitment of its activation functions to the *MLL* target genes could play a central role in leukemogenesis.

AF10 and AF17 both contain triplets of PHD fingers at their N-terminus and leucine zipper domains at their C-terminus. As discussed in chapter IV.3.1, the *MLL*-AF10 and *MLL*-AF17 fusions are invariably devoid of the PHD fingers, but contain the leucine zipper domains which are likely to mediate dimerization, which could in these cases be the key mechanism leading to transformation.

What does the diversity of fusion partners mean in terms of *MLL* mediated leukemogenesis? Currently, we are still far away from the answer, but the following considerations can constitute a valuable framework.

MLL leukemogenesis is clearly a very complex phenomenon, which has usually been framed along two antithetic models, one holding *MLL* truncation as the key pathogenic event (the "loose cannon" model), and the other emphasising the paramount importance of the fusion proteins ("gain-of-function model"). The "loose cannon" hypothesis is supported by the following observations:

1) The great diversity of unrelated translocation partners, interpreted to suggest that interruption of *MLL* is the only necessary event.

2) Aberrations which affect only *MLL* (like partial tandem duplications or deletions) have been associated to leukemias and myelodysplastic syndromes.

However, while animal studies have demonstrated haploinsufficiency for *MLL*, this did not result in a predisposition to malignancies of any type. Importantly, the *Mll*-LacZ mouse model described below (chapter) strongly argues that truncation of *MLL* per se is not sufficient, and that the contribution of other domains (in the most minimalistic model the tetramerization interface of LacZ) appears to be necessary. This obviously does not rule out

that the deregulation in hematopoietic differentiation caused by *MLL* haploinsufficiency could constitute a particularly receptive soil for additional oncogenic hits.

In contrast, the role of MLL fusion proteins in gain-of-function scenarios has found support mainly in the following findings:

1) All translocations detected in leukemias produce in-frame fusions, indicating that it is the creation of a novel aberrant protein, and not disruption of the *MLL* locus itself, which is most often selected during the rise of a leukemogenic clone.

2) A variety of cell culture and especially animal studies have demonstrated a crucial role for at least some of the fusion partners (mostly AF9 and AF19)

I would like to propose that neither the "loose cannon" nor the "gain-of-function" models can account for the full spectrum of *MLL* aberrations, and the quest for a unified leukemogenic mechanism is possibly misguided. Rather, in each specific leukemia, both the loss of one wild type *MLL* allele and the functions provided by the fusion partner could converge in different ways to deregulate cell growth. Such a "combined model" of *MLL* leukemogenesis must also incorporate the notion that some (possibly all?) fusion proteins could also exert a dominant negative function on the wild type copy of the protein, and the extent of this effect can be different for various translocations, reflecting the distinct composition in protein domains. Within this conceptual framework, the following three common themes can be recognised.

The aminoterminal portion of MLL is absolutely essential, since it is constantly retained. The simultaneous presence of the reciprocal fusion product, which would harbour the PHD and SET domains of MLL carboxyterminus fused to the aminoterminal moieties of the translocation partners, has not always been detected, arguing that it may not be necessary to the process, though it might certainly contribute additional features.

Transcriptional activation, either in the form of classical transactivation, or as enzymatic activities which alter chromatin structure, emerges as one possible domain of regulation in which some MLL fusion proteins could be involved (AF4, CBP, AF9 and AF19). However, mere transcriptional activation does not seem to be the key event, since fusions of MLL with VP16, unlike MLL-AF19, were unable to transform myeloid cells in vitro, pointing to a much more specific and finer mechanism through which fusion partners contribute to the abnormal activation of distinct gene programmes (Dimartino and Cleary, 1999; Slany et al., 1998). However, as an important corollary, transactivation of different sets of genes could contribute to the various leukemic phenotypes observed.

Aberrant MLL homodimerization through the carboxyterminal fusion partner domains is also likely to play a role. While possible in theory for every fusion, it seems particularly convincing for AF10 and AF17 (due to their leucine zipper domains) and for the partners normally located in the cytoplasm. It is also in agreement with the *Mll*-LacZ mouse model.

Finally, another convergent point of regulation might be the control of cell death. Clinically, many of the *MLL* leukemias display hyperleukocytosis, tissue infiltration and resistance to chemotherapy, all findings compatible with decreased sensitivity to apoptosis. This was observed also in t(4;11) cell lines, which displayed prolonged survival upon serum starvation when compared to other leukemic cell lines without *MLL* rearrangements (Kersey et al., 1998). While cell line experiments do not necessarily mirror the in vivo situation, other lines of research have also started to uncover the effect of *MLL* fusions in apoptosis control. *MLL*-ELL was shown to interact in vitro with p53 and to inhibit transcription from p53 responsive promoters (Maki et al., 1999). Similarly, the N-terminal part of *MLL* was

shown to interact with GADD34, which promotes apoptosis following radiation damage (Adler et al., 1999).

III.2 The t(4;11)(q21;q23) translocation

The t(4;11)(q21;q23) translocation fuses the *MLL* and the *AF4* genes. It is the most frequent translocation involving the *MLL* gene, accounting for up to 40% of all *MLL* rearrangements (Johansson et al., 1998).

The t(4;11) translocation is strongly associated with acute lymphoid leukemias (95% of 183 *MLL-AF4+* cases in the cohort mentioned above), although it also occurs with much lower frequency in acute myeloid leukemias (AMLs), acute undifferentiated leukemias (AULs), acute biphenotypic leukemias (ABLs) and treatment-related acute leukemias or myelodysplastic syndromes (sAL/MDSs). It is particularly frequent in infant acute lymphoblastic leukemias (iALLs, defined as arising before one year of age), occurring in 60% of all infant ALL cases. The frequency of this aberration is lower for other age groups of ALL patients (2% in children and 3% to 6% in adults).

Numerous studies have clearly correlated this translocation with a particularly ominous prognosis, both in children and adults, and screening for *MLL* rearrangements, particularly the *MLL-AF4* fusion, is now widely used to identify patients at high risk, who receive intensified therapy regimens. Nonetheless, so far results have been largely disappointing, with lack of improvement in overall survival or event free survival (EFS) after bone marrow transplantation in a cohort of 183 t(4;11) patients (Johansson et al., 1998). However, among t(4;11) cases, different risk stratifications can be made according to the age of the patient and the leukemic phenotype. For example, children between 1 and 10 years of

age have a longer extended event free survival (EFS) time than infants, older children or adults (Johansson et al., 1998; Pui, 2000; Pui et al., 1994; Pui et al., 1991).

In the vast majority of cases, the leukemic phenotype is lymphoblastic or mixed lymphoblastic-monocytic. Blasts are usually CD19+, CD24+, TdT+, and HLA-DR+, while they lack CD10 (the common antigen of acute lymphoblastic leukemia) and sIgM, which altogether defines a B-precursor immunophenotype. The presence of myeloid associated markers is also frequent, especially CDw65, CD13, CD15 and CD33, suggesting that the transformation event might hit a relatively undifferentiated progenitor which then proceeds up to a stage of pro-B cell block. In a series of 46 patients, there was indeed significant association between *MLL-AF4* translocation and coexpression of the myeloid marker CDw65, as compared to patients with pro-B ALL lacking the *MLL-AF4* translocation (Janssen et al., 1994).

The *MLL-AF4* rearrangement is also associated with hyperleukocytosis, organomegaly (resulting from blast infiltration) and involvement of the central nervous system (CNS), all factors which contribute to its ominous prognosis.

One of the obvious features of infant leukemias is their short latency. Two lines of evidence have established that *MLL-AF4* translocation can occur *in utero*. First, studies on identical twins with concordant leukemias demonstrated that the *MLL* gene rearrangement was nonconstitutive and hence had to have been transmitted from one twin to the other in utero via intraplacental anastomoses (Ford et al., 1993). Second, in three cases, PCR-based methods detected *MLL-AF4* genomic fusions in neonatal blood spots of individuals who developed ALL between five months and two years of age (Gale et al., 1997). Taken together, these data argue that the *MLL-AF4* fusion may be sufficient for leukemia development, with no or minimal requirement for additional genetic lesions which should

presumably result in longer latencies. For example, in the case of identical twins with the *TEL-AML1* translocation, leukemias developed with a substantially longer latency, at 3.5 and 5 years of age respectively. An even more extreme example concerns a pair of identical twins who developed T-cell leukemia at 9 and 10 years of age. Thus, at least for some leukemias, fetal origin can be coupled with very protracted latency, and therefore the strikingly rapid onset of the *MLL-AF4* disease indicates a simple molecular explanation.

On the other hand, diverse lines of research have explored the necessity of secondary oncogenic hits, both for *MLL* rearrangements in general, and for the *MLL-AF4* case in particular. First, even pediatric patients with *MLL* translocations (arguably the ones with the shortest latency), often display at the time of diagnosis additional genetic lesions, whose role in the progression of the disease is however still poorly characterised. Second, some studies reported the presence of *MLL-AF4* fusion transcripts in children affected by leukemias which were negative for *MLL* rearrangements upon Southern blot analysis, as well as, even more strikingly, in a proportion of samples from apparently healthy individuals (Uckun et al., 1998). Detection of the fusion cDNA was achieved by nested RT-PCR, attesting to the very high sensitivity needed. Contrary to ALL patients with either cytogenetic or molecular (Southern blot detection) evidence of *MLL-AF4* rearrangement, patients positive only by nested RT-PCR fared just as well as patients without *MLL* rearrangements. Taken together, these observations could mean that the *MLL-AF4* transcript was present only in a very small subset of cells (below the 1% to 5% threshold of Southern blot detection) which did not contribute further to leukemia development. In this case it still remains to be explained why these cells were not detectable anymore after chemotherapy induced remission. The same could be true for the nested RT-PCR positive cases among apparently healthy individuals (including bone marrow and liver samples from both fetuses and infants), which suggested

that this translocation could be a relatively frequent event in utero, occurring in up to 25% of normal newborns. The logical conclusion would hold that the *MLL-AF4* product is not sufficient to promote leukemogenesis. These reports were conceptually not new and added the *MLL-AF4* aberration to a growing list of genetic rearrangements which appear in a variable percentage of healthy individuals, including the t(14;18), t(9;22) and t(8;14) translocations, as well as *MLL* tandem duplication (Biernaux et al., 1995; Dolken et al., 1996; Hunger and Cleary, 1998).

However, two other studies failed to identify *MLL-AF4* transcripts in normal samples with an equivalent nested RT-PCR assay (Kim-Rouille et al., 1999). Analysis of the available data is further confounded by the different experimental strategies taken by different groups, which hinders direct comparison of the results. Under these circumstances, this issue remains hotly debated, and warrants further examination with alternative methods which could validate the nested RT-PCR results, like FISH or genomic single step PCR. Still, even if *MLL-AF4* detection was confirmed among otherwise healthy subjects, this need not directly indicate the need for multiple oncogenic hits. Rather, the whole body of information concerning *MLL* translocations indicates that these leukemias do exhibit a remarkably short latency in at least two settings, namely infant leukemias (where the *MLL-AF4* fusion is predominant) and treatment related leukemias This is in sharp contrast with most other leukemias, where the available information points to latencies of years to decades. One possibility is of course that this fusion protein can per se accelerate the accumulation of further genetic lesions. Alternatively, the sporadic occurrence of *MLL-AF4* positive cells in normal individuals would be most consistent with a model in which the fusion protein can promote oncogenesis only during a limited time window in hematopoietic development. Thus, as soon as the translocation occurs in a permissive compartment,

progression to a leukemic clone would be both very likely and fast, in agreement with the high concordance rate of *MLL* leukemias among identical twins with a monochorionic placenta, where it is clear that the leukemic clone from one twin has rapidly colonised the other. On the other hand, if the rearrangement occurred in a non conducive cellular environment, the result could well be the persistence of an otherwise innocuous clone in a healthy individual.

This brings us to the related issue of the molecular mechanisms responsible for the *MLL-AF4* translocation. Initial studies had hypothesised the improper involvement of the V(D)J recombinase normally active in the immune system, in analogy to what has been shown with translocations which fuse a variety of genes to either the immunoglobulin or the T-cell receptor gene clusters. This was largely based on sequence analysis from the 4;11 breakpoints of two cell lines (MV4;11 and RS4;11), which had detected in the immediate vicinity of the breakpoints the presence of heptamers and/or nonamers highly homologous to the recognition target sites of V(D)J recombinase (Gu et al., 1992a).

However, a recent analysis of biopsy material from fourteen t(4;11) leukemia patients strongly challenges the V(D)J recombination hypothesis (Gillert et al., 1999; Reichel et al., 1998). In all biopsies, extensive sequence characterisation of the breakpoints demonstrated deletions, duplications and inversions accompanying the translocation event. Filler DNA and mini-direct repeats were also identified at the breakpoint junctions. Thus, at the molecular level, these translocations are reciprocal but not balanced, since net loss or gain of genetic material does occur. Moreover, in this study (Gillert et al., 1999), neither target sites for V(D)J recombinase nor Alu elements (which occur in the *MLL* gene and had been previously implicated in aberrant recombination) were observed at or near the breakpoints. Overall, these findings are more compatible with this translocation being the

aberrant result of DNA repair through the so-called error prone repair mechanism (EPR), which incorporates features of the non homologous end-joining (NHEJ) pathway. This led the authors to predict that multiple DNA breaks occur in the two genes before the actual translocation actually takes place.

A potentially interesting observation which also emerged from this study was a speculation regarding a symmetrical relationship between the breakpoints on chromosome 11 and 4. The closer the chromosome 11 breakpoint was located to the centromeric end of *MLL*, the more often it was fused with the telomeric side of *AF4* on chromosome 4 and viceversa. A greater number of samples is required to strengthen this observation; however, it could reflect rather precise requirements in the relative spatial orientation of the two chromosomes during the translocation reaction, possibly connected to a reciprocal position of the respective chromosomal territories.

The "DNA repair" hypothesis could be a useful framework with which to interpret another set of studies which compared the breakpoints of de novo versus treatment related leukemias harbouring *MLL* rearrangements. *MLL* translocations occur frequently in therapy-related leukemias, after treatment with topoisomeraseII inhibitors. One function of Topoisomerase II is to unwind the DNA by introducing double strand breaks. By preventing religation of these breaks, some of the topoisomeraseII inhibitors pose a relatively high risk of translocations. An in vivo topo II cleavage site maps near exon 9 in the *MLL* breakpoint cluster region, and 11 additional sequence stretches closely related to topoII consensus binding sites are located in the telomeric portion of the *MLL* BCR (Aplan et al., 1996; Strissel et al., 1998). Moreover, detailed analysis of the BCR has revealed an asymmetric distribution of the breaks among different *MLL* leukemias (Cimino et al., 1997; Domer et al., 1995; Strissel et al., 1996). Thus, in 75% of therapy related, but only 25% of de novo

leukemias, the breakage maps to the 3' end of the BCR, in a region which has been interpreted as a scaffold attachment region (SAR). Interestingly, infant leukemias show the same bias in breakpoint distribution as therapy related leukemias, pointing to a possible common mechanism (Cimino et al., 1997). This in turn has led to the suggestion that during pregnancy many known naturally occurring topoisomerase II inhibitors, especially flavonoids, could be responsible for this translocation, an intriguing hypothesis which requires experimental tests. One clinical study has detected an association between maternal exposure to foods containing various topoII inhibitors and infant AML, but, curiously, not ALL (Ross et al., 1996). Recent in vivo studies, utilising progenitor hematopoietic cells, found that several bioflavonoids can induce site-specific cleavage in the *MLL* BCR, and that these sites colocalise with the ones induced by etoposide and doxorubicine, two known TopoII inhibitors (Strick et al., 2000).

As far as the molecular mechanisms of *MLL-AF4* leukemogenesis are concerned, very little is known, and it is limited to the results coming from several studies which investigated the transcriptional regulatory potential of *AF4* in a heterologous cell culture system. No animal model of any kind has been developed for the *MLL-AF4* translocation.

Three different studies identified a domain of transcriptional activation in the *AF4* gene by fusing different portions of the AF4 protein to the GAL4 DNA binding domain and assessing the expression of a reporter construct in a variety of cell lines. Although they utilised slightly different fusions, the results are largely overlapping, with the transactivation domain mapped respectively to residues 365-572, 480-560 and 347-475 (Ma and Staudt, 1996; Morrissey et al., 1997; Prasad et al., 1995). The region between aminoacids 458 and 572 is particularly rich in serine, proline and acidic aminoacids, which are collectively

present in the activation domains of other transcriptional regulators and thus may contribute to the activity of this domain in AF4. This putative transactivation domain is invariably retained in all MLL-AF4 fusion proteins described to date, and could contribute an important function to the fusion protein.

As expected from similar experiments with other proteins, the transactivation potential differed dramatically among different cell lines (Morrissey et al., 1997). Specifically, this domain activated transcription in COS-7 cells, HeLa carcinoma cells, NIH3T3 and MCF-7 (a breast tumor cell line), but surprisingly no transactivation was observed either in normal B cells or in B cell leukemia cell lines. In another study, FLAG-tagged MLL-AF4 fusion was expressed in U937 myeloid leukemia cells under the control of a tetracycline responsive promoter. Upon induction, MLL-AF4 caused cell-cycle arrest in G_0-G_1. Interestingly, this effect was not observed in U937 clones expressing only an MLL truncation (Caslini et al., 1996).

These results provide some evidence for a possible function of *AF4* in transcriptional regulation. At the same time, they highlight the limitations inherent to such heterologous cell systems, where the function of the fusion protein or the fusion partner is analysed on the background of a differentiation programme and/or of a genomic make-up (in terms of additional genetic lesions) which are likely to be fundamentally different from the target cell *in vivo*. However, some general conclusions can be drawn. For example, the observations in U937 cells are analogous to the results of MLL-AF19 transduction of hematopoietic progenitor cells (Lavau et al., 1997)(see chapter IV.8.2), where only low levels of transcript were detected. Plus, the same authors did not manage to express the MLL-AF19 fusion from a constitutive promoter in a variety of cell lines. Thus, a pattern is possibly emerging, whereby at least some of the MLL fusion proteins are tolerated at relatively low levels in

many cell types, causing either toxicity or growth arrest when their expression is forced at higher levels or in non permissive environments. This provides one more indication for the importance of the cellular compartment ("the soil") in leukemogenesis.

Finally, a quantitative RT-PCR analysis recently indicated that *MEIS1* and *HOXA9* are consistently upregulated in the majority of t(4;11) PrePreB ALLs. While they are also expressed in most AMLs, they are only rarely expressed in ALLs, including the PrePreB types, which lack *MLL* rearrangements. *HOXA10* had a similar expression profile in t(4;11) leukemias, but it was also expressed in most ALLs tested regardless of the underlying translocation. It is too early to conclude whether the MLL-AF4 fusion, and possibly other MLL fusions as well, have a direct role in driving expression of these genes. However, in view of the following observations, and in the light of the known function of *MLL* in *HOXA9* regulation, this result could point to a possible mechanism of oncogenesis. In fact, *HOXA9* and *MEIS1* cooperate in inducing AML, as shown both by spontaneous AMLs in BXH-2 mice and by transplanatation experiments, in which the overexpression of both genes was necessary for the development of the disease (Kroon et al., 1998; Nakamura et al., 1996b). The role of *HOXA9* deregulation in leukemogenesis is also indicated by its involvement in the t(7;11) translocation, which fuses it to the nucleoporin *NUP98* (Kroon et al., 2001; Nakamura et al., 1996a), and by the highly significant finding that its expression in human AMLs correlates with treatment failure (Golub et al., 1999). In a similar fashion, *HOXA10* is also expressed in many human AMLs, and its overexpression in mouse bone marrow cells results in AMLs with a long latency (Lawrence et al., 1999; Thorsteinsdottir et al., 1997). The molecular basis for this cooperation in leukemogenesis most likely resides in

the formation of ternary complexes containing respectively HOXA9, MEIS1 and PBX or HOXA10, MEIS1 and PBX (Schnabel et al., 2000; Shen et al., 1999).

IV

The *MLL* gene

IV.1. Introductory remarks

The *MLL* gene (for Mixed Lineage Leukemia) was cloned by four independent labs in 1992 by positional cloning of the translocation breakpoint 11q23, one of the most frequent chromosomal rearrangements in acute leukaemias (Djabali et al., 1992; Gu et al., 1992b; Tkachuk et al., 1992; Ziemin-van der Poel et al., 1991). Sequence analysis identified it as the first mammalian homologue of the *Drososphila* gene *trithorax.*

It is a large gene, spanning about 100 kb and coding for a 431 kDa protein. Diagrams of the human MLL gene and protein are shown in figure 1 and 2 respectively.

The homology to *Drosophila trithorax* is clustered in the following domains: a stretch of zinc fingers of the plant homeo domain (PHD) type, located between residues 1433-1627, and the C-terminal SET domain, also found in many other chromatin proteins. Two other regions of potential functional relevance, which are present in MLL but absent from *Drosophila trithorax*, include an N-terminal stretch of AT hooks and downstream of them a cysteine-rich domain with homology to a non-catalytic part of DNA methyltransferase. In addition, other regions of partial homology to known genes have been identified, but their significance is unknown.

It appears likely that MLL, by virtue of these domains, behaves as a modular integrator of different functions relating to chromatin dynamics. It is therefore interesting to review the most relevant findings for each of these domains, especially focusing on the possible relevance of their retention in or exclusion from the MLL fusion proteins which play a key role in leukemogenesis.

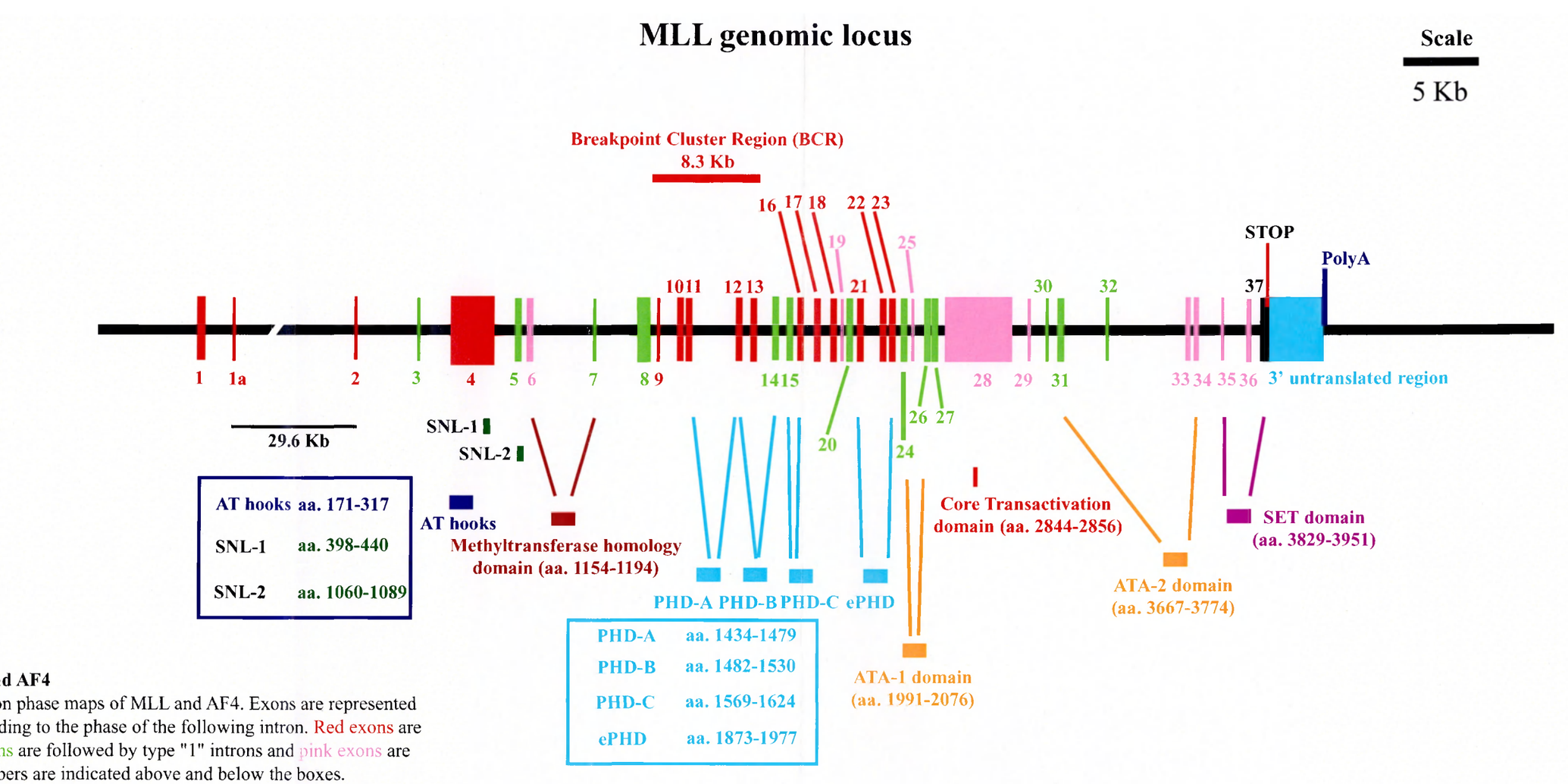


Figure 1 Intron phase map of MLL and AF4

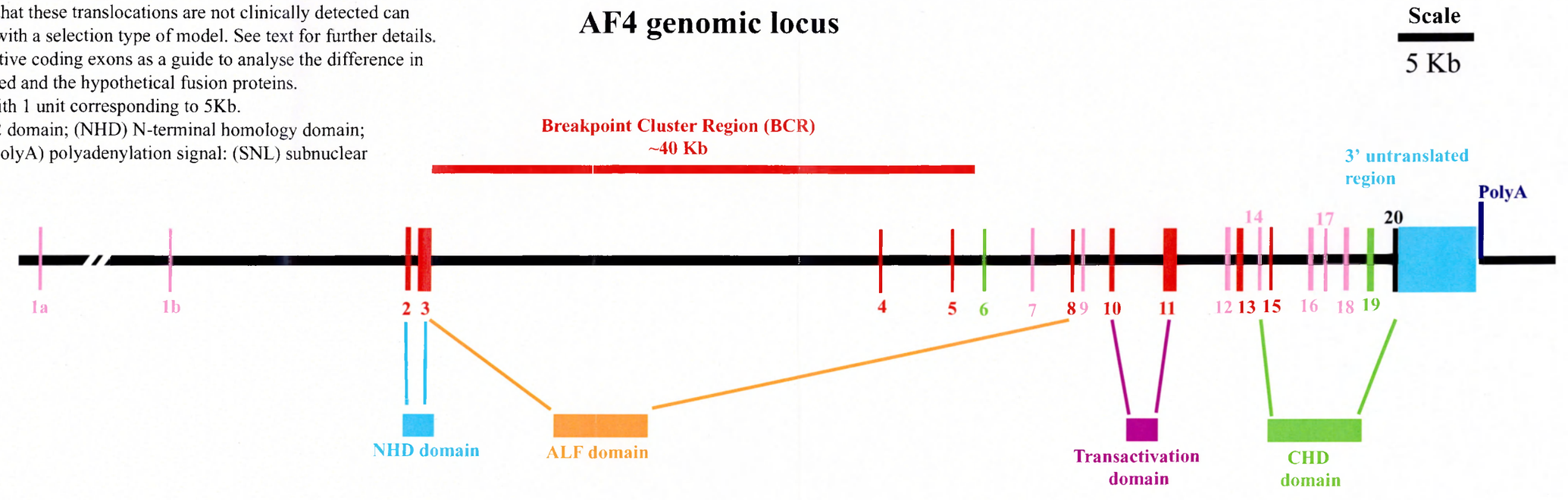
The figure shows an analysis of the intron phase maps of MLL and AF4. Exons are represented by filled boxes of different colors, according to the phase of the following intron. Red exons are followed by type "0" introns. Green exons are followed by type "1" introns and pink exons are followed by type "2" introns. Exon numbers are indicated above and below the boxes.

For both genes, the breakpoint cluster region involves only few exons. However, based purely on the intron phase, a much greater number of rearrangements could be expected which would also result in in-frame fusions. The fact that these translocations are not clinically detected can be explained either with a "hot spot" or with a selection type of model. See text for further details.

The domains are aligned with the respective coding exons as a guide to analyse the difference in domain composition between the observed and the hypothetical fusion proteins.

Both genomic loci are drawn to scale, with 1 unit corresponding to 5Kb.

Abbreviations: (ALF) AF4-LAF4-FMR2 domain; (NHD) N-terminal homology domain; (CHD) C-terminal homology domain; (polyA) polyadenylation signal: (SNL) subnuclear localization signal.



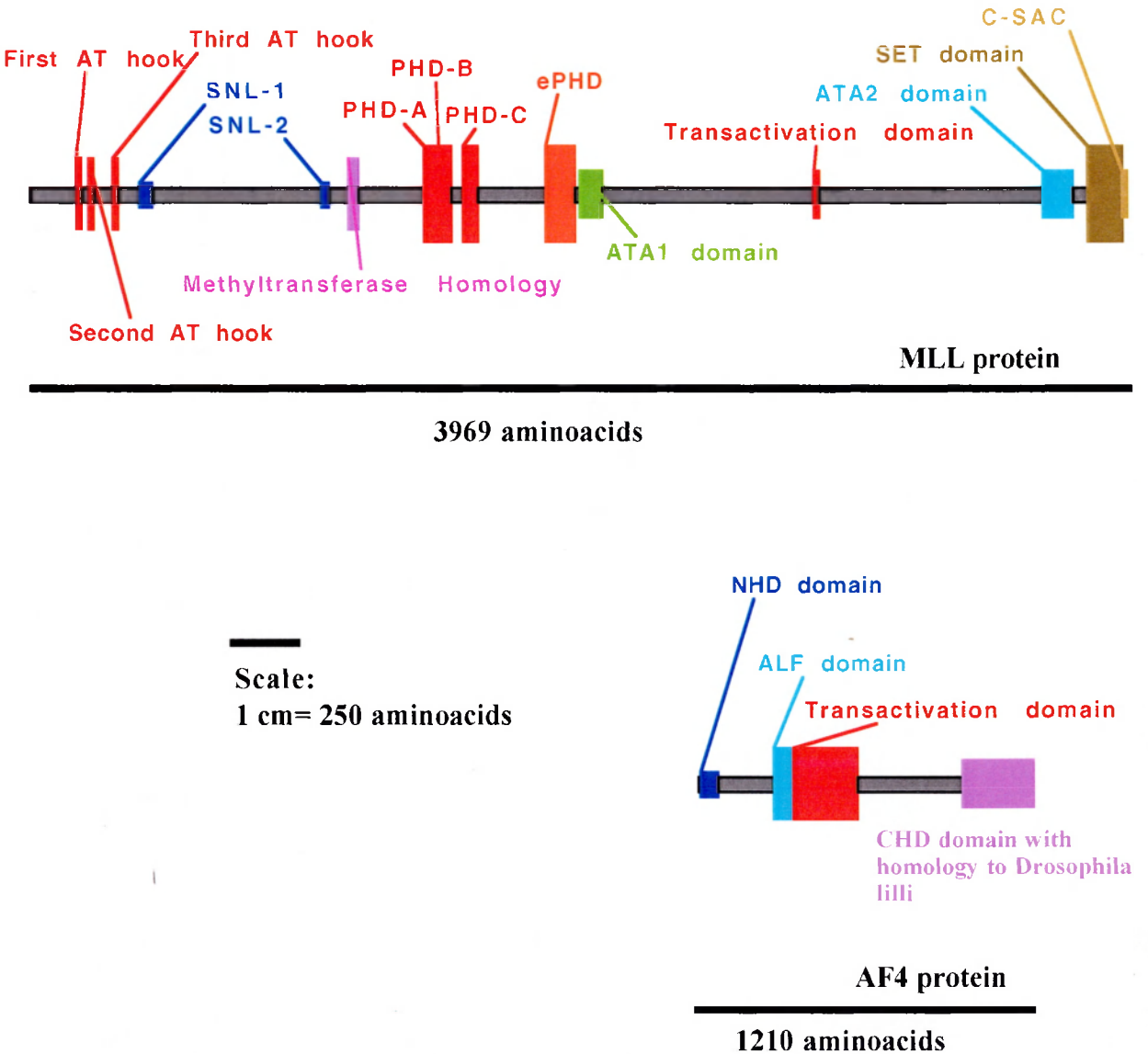


Figure 2 Diagram of the MLL and AF4 proteins
Schematic representation of the human MLL and AF4 proteins, highlighting the functional domains. The proteins are drawn to scale, with one centimeter corresponding to 250 aminoacids. For the AF4 protein, all domains, with the exception of the transactivation domain which has been functionally defined, have been assigned on the basis of homology between the AF4, LAF4 and FMR2 proteins.
Abbreviations: (NHD) N-terminal homology domain; (CHD) C-terminal homology domain; (ALF) AF4-LAF4-FMR2 domain; (C-SAC) C-terminal SET domain adjacent cysteine-rich region; (SNL) subnuclear localization signal.

IV.2 The aminoterminal domains of MLL retained in the fusion proteins

IV.2.1 The AT hooks

AT hooks were first described as a DNA binding motif in the high mobility group protein HMG-I(Y), a chromatin protein with high affinity for the minor groove of AT rich DNA stretches (Huth et al., 1997; Reeves and Nissen, 1990). They have since then been identified in many chromatin proteins. The AT hook is both necessary and sufficient for DNA binding, but, contrary to classical DNA binding motifs, AT hooks are thought to recognise a DNA structure rather than a specific nucleotide sequence. Upon binding, they can bend and distort DNA, and therefore it is hypothesised that they can regulate the architecture of promoter-enhancer interactions to facilitate transcription.

MLL has three AT hooks at its N-terminus. They have been shown to bind both cruciform DNA and scaffold attachment region (SAR) DNA (Zeleznik-Le et al., 1994).

Yeast two-hybrid screens have identified two potential protein partners for the aminoterminal region of MLL containing the AT hooks: SET and GADD34. As with any protein interaction detected under overexpression conditions, the *in vivo* relevance of these observations (summarised in table 2) remains to be elucidated.

Finally, a hint as to the possible relevance of the AT hooks to the leukemogenic capacity of the MLL fusion proteins, comes from the observation of chromosomal translocations involving the genes HMGI-C and HMGI-Y in a variety of mesenchymal tumors (Kazmierczak et al., 1998a; Kazmierczak et al., 1998b; Kazmierczak et al., 1999; Santulli et al., 2000). Similarly to the MLL fusions, these chimeric proteins also feature a triplet of AT hooks at their aminotermini, raising the possibility that they could impinge at least partially on a similar transforming pathway.

Table 2: Putative protein-protein interactions of the aminoterminal portion of MLL retained in the leukemogenic fusion proteins

Protein name	Protein features and function	Additional remarks	Reference describing the interaction
SET	- homologous to the nucleosome assembly protein NAP1 (Ito et al., 1996) - specific inhibitor of protein phosphatase 2A (PP2A) (Li et al., 1996)	-Upon overexpression, the aminoterminal portion of MLL, the SET and the PP2A proteins have been detected in a ternary complex. - The SET gene is translocated to the CAN gene in acute undifferentiated leukemias (von Lindern et al., 1992).	-Adler et al., 1997
GADD34	- induced by genotoxic stress (Adler et al., 1999) -the interaction was reported for the N-terminus of MLL, as well as for the fusions MLL-AF9, MLL-AF19 and MLL-ELL. -Following radiation, GADD34 overexpression increases apoptosis, an effect which is inhibited by concomitant expression of MLL fusions.	-upon overexpression, MLL-AF19, GADD34 and hSN5 (a member of the human SWI/SNF complex) were detected in a ternary complex.	-Adler et al., 1999

IV.2.2 The Methyl-Transferase Homology domain

A cysteine-rich DNA recognition motif is located in MLL between residues 1147 and 1194. It is a zinc finger domain of the type "CXXC", which contains eight conserved cystein residues that bind to zinc. The identification of this motif in MLL sparked great interest, since this motif had previously been identified both in proteins that methylate cytosine, and in proteins that bind to methyl-cytosine. The former include DNA methyltransferase I (DNMT1), which catalyses the transfer of a methyl group to the cytosine residues of DNA (Bestor et al., 1988; Robertson and Wolffe, 2000). In DNMT1, the CXXC type zinc finger lies towards the N-terminus. Downtream of it, there are two bromo-adjacent homology domains (BAH), which have been identified in a diverse range of proteins involved in DNA methylation, DNA replication and transcriptional regulation. Further downstream along the DNMT1 protein are three DNA methylase domains with catalytic activity. DNMT1 functions as the main maintenance methyltransferase, which copies patterns of methylation upon DNA replication. The N-terminal half of DNMT1 containing the CXXC domain is likely to exert a regulatory function on the catalytic activity of the C-terminal domains. In fact, while the intact enzyme displays little activity on unmethylated DNA, upon proteolytic cleavage of the N-terminal from the C-terminal half, unmethylated DNA is methylated much more rapidly (Bestor, 1992). These initial observations suggested that N-terminal domains could recognise the methylation status of DNA, thereby restricting enzyme activity to hemimethylated substrates. Thus, it is possible to envision that also MLL, by virtue of this CXXC domain, could recognise methylation patterns, and that this could be an important aspect of its function as a chromatin regulator. While this hypothesis still holds validity, more recent findings have provided additional insight into the function of DNMT1 and the

role of the CXXC domain. In fact, DNMT1 has been found to be associated with histone deacetylase activities in at least two different protein complexes which can repress transcription (Fuks et al., 2000; Robertson et al., 2000; Rountree et al., 2000), representing yet another route whereby DNA methylation and transcriptional repression can be coupled, together with the previously characterised route whereby methylated DNA is recognised by methyl-CpG binding protein-2 (MeCP2), which in turn recruits histone deacetylase activity (Jones et al., 1998; Nan et al., 1998).

The region of MLL homologous to DNMT1 was previously shown to repress transcription in cell culture assays upon overexpression of GAL4 fusions (Zeleznik-Le et al., 1994). This finding prompted the search for an analogous repressive function of the DNMT1 domain. Two DNMT1-GAL4 fusions, both within the region homologous to the MLL repressive domain, exhibited repression activity. This repression was at least partially relieved by trichostatin A (TSA), a known inhibitor of HDACs. Interestingly, this was true only for the construct which did not include the CXXC Zn finger, arguing that alternative mechanisms might account for the repressive property of this domain (Fuks et al., 2000).

Potentially interesting analogies for MLL can be drawn from an analysis of the two complexes containing DNMT1. One of the two complexes contains histone deacetylase 1 (HDAC1), the retinoblastoma tumor-suppressor protein (Rb) and the sequence specific transcriptional activator E2F1 (Robertson et al., 2000). Within this complex, DNMT1 is thought to interact directly with both HDAC1 and Rb, and the Rb interaction domain has been mapped to the N-terminus of the protein featuring the CXXC motifs. The E2F-Rb complex was already known to repress transcription at E2F responsive promoters via recruitment of HDAC1, a key feature of Rb function as an inhibitor of cell cycle progression. The identification of DNMT1 in this complex provided a novel link between DNA

methylation, histone deacetylation, and the Rb dependent transcriptional repression of specific promoters.

In the other complex, DNMT1 has been shown to directly interact with histone deacetylase 2 (HDAC2) and a newly identified transcriptional repressor called DNMT1 associated protein 1 (DMAP1) (Rountree et al., 2000). Once again, both of these interactions have been mapped to the N-terminal half of DNMT-1. Importantly, the repression mediated by DMAP-1 is independent of histone deacetylation activities. Taken together, these results point to a model for the function of the N-terminal half of DNMT1, based on its interacting proteins. Whereas a part of this region, most likely the one which does not contain the CXXC motifs, recruits HDAC1 and HDAC2, the more N-terminal domain, harbouring the CXXC fingers, could repress transcription through the association with Rb, DMAP1 and/or other as yet unidentified protein partners.

This overview of DNMT1 function has clear relevance for MLL. By analogy, it is possible that the CXXC fingers might allow MLL to sense boundaries of gene expression and establish protein complexes which ensure the faithful transmission of epigenetic states upon cell division.

As for the CXXC containing proteins which bind to 5-methyl-cytosine, the methyl-CpG binding protein 1 (PCM1, also called MBD1) contains three zinc-fingers domains of the CXXC type lying downstream of a domain (aa. 7-60) (Hendrich et al., 1999) which binds to DNA that contains one or more symmetrically methylated cytosines (CpGs) and is found in many methyl-CpG binding proteins from many species. PCM1 is part of the MeCP1 complex, and has been shown to repress transcription in a methylation dependent manner. As the CXXC fingers are the only other domains present in PCM1 besides the methyl-CpG

binding domain (MBD), it is once again likely that they mediate repression via recruitment of specific corepressors and/or histone deacetylases, as in DNMT1.

On the whole, the observation that the CXXC zinc-finger, other than in MLL and its closely related homologue MLL2, is specifically found only in proteins involved in either methylating DNA or in recognising methylated DNA, points to an intriguing connection between MLL function and the methylation status of the genome.

IV.3 The carboxyterminal domains excluded from the fusion protein

IV.3.1 The PHD Fingers

The PHD domain is a zinc-finger motif with a Cys4-His-Cys3 pattern, spanning 50-80 aminoacids. It was originally wrongly identified as a Lim finger in two closely related plant homeodomain proteins, HAT3.1 and HOX1A, hence the origin of its name from Plant Homeo Domain. Its identification as a novel and distinct protein domain came from a systematic search for sequence similarities between proteins of the *Trithorax* and *Polycomb* groups (Aasland et al., 1995). Subsequently, it became clear that this domain only appeared in nuclear proteins acting in chromatin-mediated transcriptional regulation. Since then, the list of proteins featuring a PHD domain has grown substantially, and it now comprises more than 400 proteins. Still, in spite of this wealth of information and the persistence of its correlation with chromatin proteins, the function of the PHD domain has remained remarkably elusive. It is clearly distinct from two other similar zinc-finger motifs, namely the RING, characterised by a Cys3-His-Cys4 pattern, and the LIM, which has a Cys2-His-Cys5 pattern. By analogy to these two domains, it is thought to be a protein-protein interaction platform, rather than a more classical type of DNA-binding zinc-finger. It belongs to a repertoire of chromatin domains (Bromo domain, BAH domain,

Chromodomain, Chromoshadow domain, RING finger, SET domain, SANT domain, SAND domain) which have been conserved in evolution, and shuffled in various combinations across different proteins and different species as modular interfaces for increasingly flexible regulation. More specifically, assortment of these domains in the trithorax and Polycomb groups, which act antagonistically to maintain the expression status of the genes of the Hox cluster, is suggestive. It may point to common mechanisms, by which these proteins could recognise their targets, although with a very different final transcriptional outcome.

Recently, structures of two PHD fingers have been solved (Capili et al., 2001; Pascual et al., 2000). The PHD finger present in the transcriptional corepressor KAP-1 confirms predictions made from initial alignments (Aasland et al., 1995). The PHD finger appears to be an autonomously folding domain, which chelates two zinc atoms in a cross-brace scheme. It is very similar to the RING finger, differing mainly in the hydrophobic core.

Provocative evidence for physiological interactions mediated by the PHD domain involve histone deacetylation as a common theme (Schultz et al., 2001; Zhang et al., 1998b). Mi2α and Mi2β, originally identified as the specific autoantigen of dermatomyositis, are members of the CHD family of proteins, which derives its name from the presence of three domains (Chromodomain, ATPase/helicase and DNA binding modules) (Delmas et al., 1993; Woodage et al., 1997). The two polypeptides are 80% identical, and contain two PHD fingers, two chromo domains, and one SWI-SNF2 type ATPase/helicase domain. Mi2β (also called CHD4) was identified as an integral component of the nucleosome remodelling and histone deacetylase complex (NurD), in which it contributes the nucleosome remodelling activity (Zhang et al., 1998b). The complex, purified by conventional chromatography, contains two histone deacetylases (HDAC1 and HDAC2) as well as the metastasis

associated factor (MTA-1) and two histone binding proteins, RbAp46 and RbAp48. In vitro experiments showed that Mi2β binds HDAC1 directly, but neither RbAp46 nor RbAp48, and that the PHD finger is necessary for this interaction, thus establishing a first potential function for the PHD domain.

In another study of the KAP-1 corepressor, the PHD and the Bromo domains in the carboxyterminal half of the protein were shown to constitute together an interaction platform for *in vivo* recruitment of the Mi2α polypeptide (Schultz et al., 2001). Interestingly, while Mi2β was not associated with KAP-1, Mi2α coimmunoprecipitated with HDAC1 and RbAp48, suggesting that the protein complex recruited through the PHD domain of KAP-1 has a distinct NuRD-like composition. In these experiments, the PHD and the Bromo domains appeared to behave as a functional unit to mediate transcriptional repression. Accordingly, upon overexpression in 293 cells, the interaction between KAP-1 and Mi2α was dependent on both the PHD and the Bromodomains. Interesting, albeit preliminary, data were obtained by assessing the repressive potential of chimeric proteins in which the PHD and the Bromo domains of various proteins were shuffled in different combinations. At most, these heterologous fusions achieved only partial levels of repression, suggesting that this was a specific property of the PHD and Bromo module within KAP-1. This observation has interesting counterparts in other studies trying to establish the function of chromatin domains, for example the methylase activity of the SET domain (see below), and may well have further implications for the future of chromatin research. It is becoming clear in fact that the presence of a domain may by itself not be enough to predict the specific function of this domain accurately within a particular protein. This is no great surprise, and could indeed be inferred from the great variability within most of these domain families, where key aminoacid residues provide a common scaffold, while the functional versatility lies in the

details. Hence investigations of common domains from specific proteins constitute an invaluable reference framework, but a thorough understanding of chromatin regulation will require the characterisation of each and every protein complex involved. Interpretative shortcuts should be viewed with caution.

MLL has three canonical PHD fingers, clustered between residues 1433-1627, and downstream of them one so-called extended PHD finger (ePHD). The extended PHD differs from the canonical PHD finger by having an additional Cys_2-His-Cys_4 motif at its C-terminus. All reported leukemic MLL translocations interrupt the protein upstream of (most often) or immediately at the beginning of the triple PHD stretch. The resulting fusion protein is therefore almost invariably devoid of PHD fingers. Although they would be included in the reciprocal chimeric protein (along with the C-terminally located SET domain), the reciprocal fusion protein is not detected in all leukaemias, and its pathogenic relevance is therefore still debated. As truncation of MLL eliminates the PHD and the SET domains, a model was suggested whereby this truncation would itself constitute the first event on the road to malignant transformation (Yu et al., 1995). The second event would most likely come from the domains contributed by the fusion partner and/or by other unidentified genetic alterations. There is some evidence, also from experimental animal models (see below), to at least partially support this model. Alternatively, the truncated part of MLL in the various fusion proteins could act as a dominant negative with respect to the wild type (wt) MLL and its interacting proteins.

A further hint to the possible role played by loss of PHD function in the leukemogenic process comes from the characterisation of two MLL translocations which fuse MLL respectively to AF10 or AF17. Both genes have an N-terminally located PHD finger, followed by an extended PHD finger. Remarkably, all translocations examined to date interrupt these genes downstream of the extended PHD finger, with the result that the

ensuing MLL-AF10 and MLL-AF17 fusion proteins are completely devoid of PHD fingers. Lately, the ePHD finger of AF10 was shown to mediate homo-oligomerisation *in vitro;* interestingly, homo-oligomerization was required for the ability of the AT-hook domain (also present in MLL) of AF10 to bind DNA (Linder et al., 2000). While awaiting a confirmation of these results *in vivo*, it is intriguing to hypothesise a similar function for the MLL ePHD, which could mediate either homodimerisation or heterodimerisation with other ePHD containing proteins.

On the whole, the function of the MLL PHD fingers is unknown. From the evidence gathered for the PHDs of other proteins, in MLL this domain could recruit complexes containing either Mi2α (by analogy to KAP-1) and/or HDAC1 (by analogy to Mi2β within the NuRD complex). Interaction with repressive complexes, however, need to be reconciled with the activating functions of MLL and *Drosophila* Trx that have been genetically demonstrated by *in vivo* studies. One possibility would be for MLL to act as repressor of repressors, yielding a net activation function. Alternatively, MLL might oppose repressive complexes at specific target sites to quench their activities and thus enable local maintenace of gene expression.

IV.3.2 The SET domain

The **SET** domain (for **S**u(var)3-9, **E**nhancer of zeste and **T**rithorax), evolutionarily conserved from yeast to man, consists of approximately 130 aminoacids, and is present in at least 140 proteins involved in chromatin regulation (Jenuwein et al., 1998). Like the PHD finger, the SET domain is found in very different classes of chromatin proteins, often displaying antagonistic functions in the regulation of gene expression. For example, Enhancer of Zeste (E[z]) belongs to the Pc-G group of proteins, which establish repressive chromatin domains, while trithorax (Trx) is the founding member of the trx-G group of proteins, which are required to maintain gene activity.

Multiple alignments between fifteen representative members of the SET domain family of proteins have identified within the SET domain the regions of highest sequence identity and hence presumptive functional significance (Jenuwein et al., 1998). The first region, of about 20 residues, was found to be invariant among all members analysed, and was predicted to adopt an α-helical configuration by secondary structure predictive algorithms. The second region comprises about 50 aminoacids, with an inner core forming a predicted strand-loop-strand structure, and two to three conserved histidine/cysteine residues. The spacing between these histidine/cysteine residues, together with the overall alignment results, divides SET domain proteins into four different subgroups, which have been named after their founding members, Enhancer of zeste [E(z)], Trithorax (Trx), Absent small or homeotic (Ash1) and Suppressor of variegation 3-9 [SuVar(3-9)] (reviewed in Jenuwein et al., 1998).

Very recently, both the mammalian SUVAR3-9 and Clr4 proteins were shown to be histone methyltransferases which specifically methylate lysine 9 on the amino terminus of histone H3 *in vitro* (Rea et al., 2000). This first evidence for an enzymatic activity of the

SET domain provides a starting point from which to investigate the functions of other SET domains as well. The catalytic site was mapped by mutational analysis to the motif Hϕϕ NHSC (where ϕ represents a hydrophobic residue), corresponding to residues 320-326 in the HUSUV3-9 protein. However, methylation also depended on the C-terminal tail and the cysteine-rich region, which flank the SET domain. Deletions of these elements abrogated enzymatic activity. This may explain the finding that neither MLL nor EZH2 displayed any methylase activity in these experiments, as MLL lacks the cysteine-rich region N-terminal to the SET domain, and EZH2 lacks the carboxyterminal tail.

Methylation at specific lysine residues, along with phosphorylation and acetylation on histone tails are pivotal in chromatin regulation. For example, specific modifications of the N-terminus of histone 3 at various positions have been linked to distinct chromatin functions: acetylation at lysine 14 with transcriptional activation, acetylation at lysine 9 with histone deposition and phosphorylation at serine 10 with chromosome condensation (Strahl and Allis, 2000). On the molecular level, these modifications provide docking sites for the appropriate chromatin modifiers. At least two more layers of regulation increase the flexibility and complexity of this histone information system. The first is the action of enzymes which catalyze opposing reactions in a highly specific and regulated fashion, among which histone deacetylases have been best characterised. Demethylases have been postulated but not yet identified. Second, it is becoming clear that individual modifications can affect each other. Thus, in the case of SUV39 methylase activity, it was shown that phosphorylation at serine 10 inhibited SUV3-9 mediated methylation at the adjacent lysine 9, and conversely, dimethylation at lysine 9 substantially decreased serine 10 phosphorylation by Aurora kinase (Rea et al., 2000).

Furthermore, as histones are not the only targets for histone acetyl transferases (HATs), it appears likely that methylation might also not be restricted to histones. An example comes already from *Pisum sativum*, where a lysine methylase methylates the large subunit of a metabolic enzyme.

Within this framework then, it is possible to envision that different SET domains have specificities for different residues in histone tails and/or for other nuclear targets. While substantial variation occurs within the core of the SET domains identified so far, and could thus provide the molecular basis of differential target recognition, the basis for this specificity could also come from the adjacent N-terminal region, that is critical for SUVAR3-9 catalysis on histone tails. Interestingly, MLL and related proteins have a different conserved motif in this position. Furthermore, the potential catalytic site of MLL (RϕϕNHSC) matches the same region in Clr4. This suggests that MLL is a methylase and its substrate is different from the histone substrates used by Rea et al (2000).

An interesting alternative holds that some SET domains, including that of MLL, are truly devoid of methylase activity. Rather, they might counteract the function of active methylases by binding to the same targets. This would be an analogous role as that played by anti-phophatases, and appears particularly intriguing in the case of MLL since one of its potential protein partners (Sbf-1, see below) belongs to the anti-phosphatase family (Cui et al., 1998) (see table 3). The MLL SET domain could thus conceivably counteract two enzymatic activities and their related pathways at the same time.

The relevance of this SET domain function is supported by recent experiments in which histone 3 methylation at lysine 9, but not at lysine 4, created a binding site for the chromodomain of the heterochromatin protein HP1 (Lachner et al., 2001).

Of the protein interactions described for the MLL SET domain (summarised in table 3), all were originally identified by yeast-two hybrid assays, and confirmed in heterologous cell systems with overexpression studies. Their specificity and significance await further confirmation.

IV.3.3 The ATA1 and ATA2 domains

The initial comparison between MLL and Trx drew attention to additional, shorter regions of shared homology (Stassen et al., 1995; Tillib et al., 1995). Two of them were termed ATA1 and ATA2. In both MLL and Trx, the ATA1 domain is located downstream of the ePHD finger, while the ATA2 motif immediately precedes the SET domain.

Database searches indicate that the ATA1 domain is found only in association with the ePHD in certain cases. It features several conserved hydrophobic residues and is predicted to contain four to five βstrands and one αhelix.

The ATA2 domain is estimated to be 107-148 residues long, and to contain two αhelices and several βstrands. This motif flanks the SET domain in all Trithorax homologs and occupies the immediately adjacent N-terminal position taken by the cysteine-rich region of Su(var)3-9 members, which appears to be crucial for the methylase activity.

Table 3: Putative protein-protein interactions of the SET domain of MLL, which is excluded from the leukemogenic fusion proteins:

Protein name	Protein features and function	Reference describing the interaction
Sbf-1 (SET binding factor-1)	-homologous to myotubularin, a dual specificity phosphatase, but Sbf-1 lacks this activity.	Cui et al., 1998
INI1	-member of the SWI/SNF complex -Mutations of INI1 have been linked to pediatric rhabdoid tumours and to lymphoid malignancies (Versteege et al., 1998; Yuge et al., 2000)	Rozenblatt-Rosen et al., 1998
MLL SET domain	-no function ahs yet been assigned to the MLL SET domain. In an in vitro assay for methylation at lysine 9 of the histone H3 tails, the MLL SET domain, in contrast to the SET domains of the SUVAR3-9 and CLR4 proteins, did not show any activity (Rea et al., 2000)	Rozovskaia et al., 2000

IV.4 The *MLL* genomic locus: hot spots versus selection

The *MLL* gene in humans and mouse is greater than 100kb, and has been shown to contain at least 37 exons, updating previous reports which had estimated only 21 exons (Nilson et al., 1996). The new nomenclature is used here. The breakpoint cluster region (BCR) has been completely sequenced; it is 8.3 kb long and covers the region from the 3' end of exon 8 to the 5' end of exon 14 (Gu et al., 1994; Marschalek et al., 1995). Except for intron 8, which is of type 1 (ie. interrupting the codon after the first base), all introns within the BCR are of type "0" (ie. they interrupt the coding sequence between triplets) (figure 1). The same is true for the *AF4* BCR (Nilson et al., 1997), providing the basis for the creation of in-frame fusion proteins following translocation. The narrowness of the *MLL* BCR, which contrasts with the much wider breakpoints observed in other translocations, is one of the most intriguing features of *MLL* translocations, and can be explained by two alternative possibilities. The first hypothesis predicts that rearrangements only occur within this region due to the presence of a "hot spot" sequence which promotes illegitimate recombination ("hot spot" hypothesis); and potentially hot spots also occur in BCRs of the *MLL* translocation partners. The alternative view holds that only translocations occurring between certain introns produce in-frame fusions, whose specific combination of protein domains confers a selective growth advantage (selection hypothesis). According to this hypothesis, all combinations which have not been clinically reported probably occur with similar frequency, but result either in out-of-frame products or in in-frame fusions with negligible effects.

Intron phase maps of the *MLL* and *AF4* genes are shown in figure 1. The BCR of *AF4* spans at least 40 kb and involves three type "0" introns (3, 4 and more rarely 5). The involvement of intron 2, which is also of type "0", is uncertain, since most published studies used primer sets which only explore the region downstream of exon 3. However, at least one

study reported mapping of breakpoints which presumably occur upstream of exon 3 (Chen et al., 1993).

The tight clustering of *MLL* breakpoints lends strong support to the selection model. It suggests that the aminoterminal domains of MLL (AT-hooks, nuclear localisation signals and methyltransferase homology domain) are necessary for the oncogenic function of the fusion proteins. Concomitant exclusion of the downstream domains (PHD fingers, ATA1, ATA2 and SET domains) could be equally relevant.

Following this hypothesis, the requirement for the *MLL* break to occur in this specific region of the protein directs the translocations to a patch of type "0" introns, many of which can be apparently used. Concomitantly, this directs fusions to type "0" introns of the translocation partners, including *AF4*. Interestingly, there is a single reported exception to the involvement of type "0" introns in the *MLL* BCR. In this case, the type "1" intron 8 is translocated to a presumptive type "1" intron of AF10 (Chaplin et al., 1995b). Since intron 8 of *MLL* is immediately adjacent to the type "0" introns of the *MLL* BCR, this maps the border of the required protein sequence for leukemogenesis by MLL fusion proteins to exon 8. No intron phase maps are available for other translocations partners, but the detection of in-frame fusions in all cases necessarily implies that the breakpoint regions of these genes occur in type "0" introns, except for the one case mentioned above.

If one considers the overall distribution of intron types in *MLL* and *AF4*, it is clear that many more possibile rearrangements exist which would also give rise to in-frame fusions, through other introns of the same type. The fact that essentially only those directed by the intron "0" phase of *MLL* have been observed is strong evidence for the selection model, which also predicts that the rearrangements are not detected because they do not result in a growth advantage.

MLL-AF10 translocations provide an interesting case. *MLL* and *AF10* have opposite telomere-centromere orientations with respect to the direction of transcription, and constitute the first example of oppositely oriented genes involved in translocations which yield in-frame fusions. This is possible because they undergo complex rearrangements, sometimes involving even three chromosomes, in which one of the two genes is first inverted and then translocated (Beverloo et al., 1995; Chaplin et al., 2001; Tanabe et al., 1996). Two homologues of *AF10* have been described in mammals, *AF17* and *BR140*, and *AF17* is also translocated to *MLL* (Chaplin et al., 1995a; Gregorini et al., 1996; Prasad et al., 1994). These proteins have a similar architecture consisting of PHD fingers located towards the N-terminus and leucine zippers positioned at the C-terminus. Notably, all breakpoints mapped to date for *AF10 and AF17* occur between these two domains, suggesting that their respective exclusion from and retention in the MLL fusion proteins are essential for leukemogenesis (Chaplin et al., 1995b).

Although AF10, AF17 and BR140 are extremely similar (93% and 77% identity respectively in the PHD finger and the leucine zipper between AF10 and AF17, and 66% similarity between AF10 and BR140 over a region of 150 residues), AF17 is much less frequently involved than AF10, and BR140 has not been reported in *MLL* translocations, in spite of the fact that *MLL*-AF10 translocations require an additional chromosomal rearrangement to produce a fusion protein. This is very strong evidence for the selection hypothesis because it indicates that chromosomal rearrangements are not rate limiting. Rather, it suggests that in spite of the healthy homologies between AF10, AF17 and BR140, *MLL*-AF10 fusion proteins have the greater oncogenic potential.

The *AF10/AF17/BR140* case sheds light on other cases of differential *MLL* translocation frequencies to homologous partners. They include translocations of *MLL* to *AF4* and *AF5*q31, *CBP* and *p300*, and *AF9* and *ENL*. In all pairs, the more frequent partner is listed

first. Since all translocations have been found in leukemias, the very different frequencies provoke the attractive possibility that the more frequent events are promoted by hot spots. This conclusion, however, assumes that the oncogenic potential of each pair is identical.

CBP and p300a are highly homologous transcription coactivators; however, while fusion with CBP has been found in several cases (Rowley et al., 1997; Satake et al., 1997; Sobulo et al., 1997; Taki et al., 1997), so far only one case has been reported of *MLL-p300* translocation (Ida et al., 1997). In MLL-CBP fusions, virtually whole of the CBP protein is retained, except for the nuclear hormone receptor interaction domain (NID). In contrast, in the *MLL*-p300 case, the breakpoint is further towards the C-terminus and the fusion product lacks both the cysteine-histidine rich region and the CREB binding domain located towards the N-terminus of p300. Thus, the strikingly different frequency with which *CBP* and *p300* are involved might simply mean that the observed *MLL-p300* fusion (i) is the result of the only possible in-frame translocation (based on the distribution of intron types) and (ii) has a lower oncogenic capacity if compared to the MLL-CBP fusion. Alternatively, these two genes, which lie on different chromosomes, could be translocated to *MLL* with different frequencies, reflecting distinct features in the primary sequence or in the accessibility of their loci. However, the assumption that they are functionally identical has been disproved (Kawasaki et al., 1998).

Similarly, *AF4* and *AF5q31* are highly homologous proteins, and also in the case of the *MLL-AF5q31* fusion, breakage occurs within the ALF domain (Nilson et al., 1997; Taki et al., 1999). However, while *AF4* is the most frequent translocation partner of *MLL*, so far the *MLL-AF5q31* translocation has been detected only in one case. The homology between *AF4* and *AF5q31* (57% in the transactivation domain) is lower than between *CBP* and *p300*, and therefore their different frequency of involvement might reflect genuine functional differences.

AF9 and *AF19* (also called *ENL*), which share 56% identity and 68% similarity with the highest homology at the amino- and the carboxitermini, are also translocated to *MLL* with different frequencies, accounting for, respectively, 27% and <12% of all *MLL* translocations in a recent cohort (Secker-Walker, 1998; Swansbury et al., 1998). Two BCRs have been defined in *AF9*, spanning respectively intron 4 and introns 7 and 8. A recent study identified in these introns both an *in vivo* topoII cleavage site and two scaffold attachment regions (SARs); as these elements are also present in the *MLL* breakpoint, it was suggested that they play a role in mediating *MLL-AF9* translocation (Strissel et al., 2000).

Finally, a close homologue of *MLL* was recently identified in the Stewart lab (Angrand, P.O. et al. unpublished results) as well as in two other laboratories, and called *MLL2* (FitzGerald and Diaz, 1999; Huntsman et al., 1999). *MLL* and *MLL2* have the same domain architecture and are clearly the product of a gene duplication event (Angrand et al., unpublished observations). Although the overall identity is 25%, the PHD finger and the SET domain regions are 63% and 74% identical respectively. Though a much smaller gene (21 kb), *MLL2* also has 37 exons. Yet, so far *MLL2* has not been detected in any chromosomal translocation, an intriguing observation which could point to a functional difference with *MLL*. On the other hand, it could simply mean that this locus has lost the "hot spots" which putatively drive *MLL* rearrangements.

Partial support to the selection hypothesis comes from the analysis of Alu elements in the *MLL* gene. These are repetitive elements of about 300 bp scattered throughout the human genome, and consist of two homologous tandem repeats separated by an AT-rich stretch. In the case of *MLL*, their presence has been invoked as the structural basis for chromosomal rearrangement through aberrant homologous recombination by unequal crossing-over during meiosis. Partial tandem duplication in *MLL* indicates at least in some cases improper homologous recombination between Alu elements (Schichman et al., 1994a; Schichman et

al., 1994b; Strout et al., 1998). This usually results in fusion of one of the BCR introns with the intron downstream of exon 2. Yet, several other *MLL* introns contain Alu repeats, and recombination between some of them could produce in-frame fusions, which so far though have not been reported. The potential contribution of Alu repeats to *MLL* rearrangemenst was recently addressed in an analysis of 45 kb of sequence spanning the first three exons, which indicated that some *MLL* duplications cannot be explained by the presence of Alu repeats (Wiedemann et al., 1999), and therefore their role in *MLL* interchromosomal translocations remains an open issue.

On the whole, the variety of *MLL* rearrangements all centered around a narrow breakpoint region presents a complex biological problem, and underscores the relevance of the aminoterminal MLL moieties in leukemogenesis. All available evidence can be most simply explained by the selection hypothesis, although features of genomic instability in the BCRs of *MLL* and its partner genes are likely to further restrict the range of possible aberrations.

IV.5 Trithorax and Polycomb group genes in flies and mammals

The homology with *Drosophila* trithorax has contributed to the fundamental model for MLL function, both in normal development and leukemogenesis. Thus, MLL action has been interpreted within the experimental and conceptual framework of trithorax group (trxG) and Polycomb group (PcG) genes respectively maintaining stable states of gene activity or repression. Here some of the main features of these two families of genes are summarised, focusing on the data which are most relevant to MLL function.

In the fly, the identity of body segments is established early in development as a readout of the relative activities of homeotic (Hox) genes, which are clustered in *Drosophila* in two complexes, the Antennapedia (ANT-C) and the Bithorax (BX-C) complex (Duncan, 1987; Kaufman et al., 1990; Kennison, 1995). Loss or gain of function of the appropriate Hox genes in the cells of a specific segment results in the formation of structures characteristic of a different body segment, a phenotype usually referred to as homeotic transformation. The appropriate patterns of Hox expression are established early on through the activity of maternally encoded factors, the pair rule and the gap genes. However, some of these transcription factors are present only transiently during development, while the correct Hox expression profile is required throughout the life of the organism Therefore, other genes have evolved to maintain these expression patterns stably, the trxG and the PcG genes.

Starting in 1940, with the identification of the *extra sex comb* (esc) mutation, the founder of the Polycomb family, many genes have been identified whose loss of function causes a "polycomb-like phenotype", that is, alteration of segmental identities as a result of particular Hox genes derepression (Simon, 1995). However, although Hox genes have been the best studied example, also other genes are stably repressed by this family of proteins.

In 1981, the trithorax mutation was identified. It was characterised by inadequate expression of certain Hox genes, leading to transformations in segment identities and displaying a maternal effect (Ingham, 1981; Ingham, 1998). It was soon recognised that this mutation suppressed the polycomb phenotype, leading to the current paradigm that these two classes of proteins act in an antagonistic fashion to maintain a repressed (PcG) or an active (trxG) state of Hox gene expression. Since then, virtually all other trxG genes have been identified through genetic screens searching for suppressors of polycomb phenotypes (Kennison and Tamkun, 1988; Kennison and Tamkun, 1992). This experimental strategy is worth emphasising, as it can account for one major difference in the general features of these two families of genes. Though extremely complex and varied, the PcG family shows substantial homogeneity, in that many members contain shared domains, have been found associated together in large multiprotein complexes, and virtually all appear to mediate transcriptional repression, albeit through a still poorly characterised mechanism (Kennison, 1995). In contrast, members of the Trithorax group appear to constitute a heterogeneous set of proteins. This is hardly surprising if one considers that affiliation to this group is granted solely upon a suppressive genetic interaction with any one of the many PcG genes identified. Hence, according to the above definition, interference at any level in the pathway mediating PcG action can uncover a trxG mutation.This in turn implies that the molecular mechanisms underlying the function of trxG proteins are likely to be significantly more diverse than for the Polycomb group.

The general relevance of this genetic paradigm in flies is emerging with the identification of mammalian homologues for many of trxG and PcG genes, of which *MLL* was one of the first examples. Importantly, for some of them, an analogous, reciprocal

function in the regulation of subsets of Hox genes has been documented (Hanson et al., 1999; Yu et al., 1995).

In terms of the molecular mechanisms of action, PcG proteins in *Drosophila* mediate gene repression through so called Polycomb response elements (PREs), which are complex and still partially characterised sequences, varying from hundreds of base pairs to a few kilobases (Paro et al., 1998; Pirrotta, 1998). They were initially defined by their capacity to maintain repression of hox genes in transgenic flies and/or repress reporters in transgenes in a Polycomb dependent manner, that is, mutations in PcG genes affect the activity of these elements. On the basis of colocalisation on polytene chromosomes, it was initially suggested that most of these proteins would function within large multimeric complexes assembled at the relevant target sites. Recent biochemical data provide strong evidence for this model, and two bona fide Polycomb complexes have been characterised, both in flies and mammals (Alkema et al., 1997; Ng et al., 2000; Shao et al., 1999; van Lohuizen et al., 1998). It is thought that different Polycomb complexes bind to different regions within a single PRE, but it remains to be determined how they are recruited to these elements. In fact, until now, only one member of the family, Pleiohomeotic (Pho), has been directly shown to bind a PRE, but interestingly no other PcG members have been found to interact with it (Brock and van Lohuizen, 2001; Brown et al., 1998).

Trithorax response elements (TREs) have been also postulated. The overall evidence suggests that, if TREs exist, they are closely intermingled with PREs (Tillib et al., 1999). For example, the GAGA factor, encoded by the trxG gene Trithorax-like (Trl), is bound *in vivo* to some PREs within the BX-C, though possibly in a different manner than PcG proteins (Farkas et al., 1994). Reflecting the multifaceted nature of these complex regulatory regions,

it was recently proposed that they be renamed maintenance elements (MEs), leaving the thorough characterisation of each of them to the task of identifying the respective binding sites for the Polycomb or Trithorax complexes (Brock and van Lohuizen, 2001).

These observations resonate with a variety of findings which have recently shown some potential weaknesses in the traditional, strictly antagonistic model of Polycomb versus Trithorax. Some Polycomb group proteins are needed to maintain activation, as in the case of Posterior sex combs (Psc), which is required for polyhomeotic (ph) expression (Fauvarque et al., 1995). Conversely,mutations in the trxG member GAGA factor cause derepression of the Fab-7 PRE, arguing that at least some trxG proteins can also mediate a repressive function (Hagstrom et al., 1997). More recently, the chromatin remodelling complex containing Osa and Brahma, two trxG members, was shown to be required for repression of wingless target genes (Collins and Treisman, 2000). Finally, the enhancer of zeste (E(z)) gene, originally classified as a PcG gene, was shown to function also as a trxG gene (LaJeunesse and Shearn, 1996), and in a genetic screen aimed at identifying enhancers of the Trx phenotype, both TrxG and PcG mutations were identified, indicating that at least some members of these two families might have a dual role in regulation of gene expression (Gildea et al., 2000).

How do these and similar observations relate to our understanding of MLL action? The Trx framework is obviously still very valid, and has gained important support from *MLL* knock-out studies which uncovered a role for MLL in maintaining transcription of at least some Hox genes, whose deregulation could also account for some oncogenic properties of *MLL* mutations, as various Hox genes have been directly implicated in leukemogenesis (Lawrence et al., 1996). However, the two domains consistently retained in all MLL fusions, AT hooks and methyltranferase homology, are notably absent from *Drosophila* Trx, arguing

that during evolution additional functions were conferred upon MLL which might not find an adequate parallel in the fly Trithorax paradigm. Furthermore, it is clear from the heterogeneity of the trx-G, and even more from the complexity inherent in Polycomb/trithorax functions, that the mere classification of *MLL* as a trxG member does not bring much informational content. Rather, as with all other trxG proteins, dedicated, *in vivo* studies of function and the interacting protein partners are needed to unravel its complex action, including the Trx aspects involved therein.

IV.6 Cell culture studies of *MLL* function

Several studies have addressed different aspects of MLL function in cell culture. They have mostly focused on one of two main topics: the potential role of MLL and some of its fusion proteins as transcriptional regulators, and its nuclear localisation.

Different regions of MLL were fused to the DNA binding domain of the yeast GAL-4 protein and assessed for the transcriptional regulation of a reporter gene harbouring GAL-4 binding sites. The experiments were conducted in HeLa cells, NIH3T3 and CHO (Chinese hamster ovary) cells. The results from two independent studies are largely overlapping (Prasad et al., 1995; Zeleznik-Le et al., 1994). A transcriptional activation domain was originally identified between aminoacids 2339 and 3759; further characterisation identified a minimal transactivation domain spanning residues 2829 through 2883, which yielded 300- to 500-fold activation of the reporter in this system. Mutational analysis identified a core sequence crucial for this activity: ILPSDIMDFVL, composed of two aspartate residues and seven hydrophobic aminoacids. Substitution of either the aspartic acid or the hydrophobic residues almost completely abolished the activity. This aminoacid array is reminiscent of similar motifs found in the activation domains of known transcriptional regulators, like VP16. One possible basis for the function of this transactivation domain comes from its

recent identification in a yeast three-hybrid screen searching for interacting partners of the CBP coactivator protein (creb binding protein) (Ernst et al., 2001). The screen was designed to isolate products which could interact with the unique interface formed by the CREB-CBP interaction. The interaction with the MLL transactivation domain was mapped to the CREB binding domain of CBP (also called KIX domain) and was confirmed by coimmunoprecipitation of the overexpressed epitope-tagged domains. Interestingly, in cell culture overexpression assays, increasing amounts of the E1A12S protein (a virally encoded polypeptide which inhibits CBP-dependent activation by a variety of transcription factors) abolished the transactivation potential of the MLL fragment (aa 2829-2883 of MLL), strongly suggesting that its transactivation requires CBP. Mutational analysis identified several residues, within the core motif described above, which appear to be crucial for both CBP interaction and transactivation. If confirmed *in vivo*, these findings could provide an attractive mechanism for the role of MLL in enabling active states of gene expression. Furthermore, the MLL-CBP fusion proteins from t(11;16) translocations could be constitutive, aberrant surrogates of the normal CBP-mediated MLL regulatory function.

Interestingly, a transcriptional repression domain was mapped to residues 1032 through 1395, which include the region of homology to methyltransferase (1147 to 1240), which independently showed the highest repression (12-fold). This domain is retained in all MLL translocation products, as well as in the cases of tandem duplication.

Several studies showed that MLL has a punctate nuclear distribution pattern, consisting of small speckles distributed throughout the nucleoplasm; the protein appeared to be excluded from the core of the nucleolus. These observations were carried out in a variety of transformed cell lines (among which COS, HeLa, SV80, Jurkat, HL60, U937), including

leukemic cell lines with 11q23 rearrangements (ML2, RS4;11 and HB11;19), as well as in normal human tissues (Yano et al., 1997).

To identify the specific motifs involved in nuclear localisation, multiple fusions were assessed in which parts of the MLL protein directed the localisation of the cytoplasmic enzyme PK (Yano et al., 1997). Two classes of elements were identified, responsible for nuclear localisation per se and/or for the specific dotted pattern, respectively. Overall, the 820 aminoterminal residues exhibited nuclear targeting properties, clustered in three motifs: NTS-1, NTS-2 and NTS-3. The latter displays a classical nuclear localisation motif of RKRKRK sequences. Two elements identified in the N-terminal portion of the protein were responsible for the speckled nuclear distribution of MLL: SNL-1 and SNL-2, respectively spanning residues 388 to 432 and 1051 to 1089). They are located between the AT hooks and the CXXC domain. SNL-1 overlaps with NTS-2. Strikingly, these motifs are conserved between MLL and *Drosophila* Trx, arguing that specific subnuclear localisation of these chromatin regulators has been strictly conserved throughout evolution. Both these localisation elements are retained in the MLL translocation products. Although the N-terminal part of MLL contains additional subnuclear targeting motifs, in this assay no elements C-terminal of the translocation breakpoint were able to direct either nuclear localisation or distribution into speckles.

Another study extended the analysis to an N-terminal MLL truncation, which retained the AT hooks and most of the SNL-1 region (Caslini et al., 2000a; Caslini et al., 2000b). This showed the same localisation pattern as the endogenous full length protein, with a punctate nuclear distribution, which included regions adjacent to the nuclear envelope and the periphery of the nucleolus. Intriguingly, different imaging approaches (confocal laser

scanning and immunoelectron microscopy) showed colocalisation of this truncated Mll protein with topoisomeraseII in metaphase and telophase during mitosis.

Interestingly, for at least some of the fusion proteins (MLL-AF4, MLL-AF9 and MLL-AF17), the subnuclear distribution was not affected (Yano et al., 1997), lending to the suggestion that in the chimeric products, the aminoterminal MLL moieties target the domains of the translocation partners to the normal, physiological targets of MLL. This in turn is predicted to result in aberrant regulation of those targets.

These studies also indicate that, to be detected in bright dots by immunofluorescence, the Mll protein must be present in multiple copies at its target sites within the nucleus. This is, at least conceptually, in agreement with the long standing notion that MLL, like most trx-G members functions in large multimeric complexes. Furthermore, the observation that MLL nuclear dots do not overlap with nuclear subdomains defined by other oncoproteins, like PML (Promyelocytic leukemia) and TAL-1 could point to a novel and distinct nuclear subdomain where MLL exerts its action .

On the whole, several approaches have convincingly demonstrated the role of the N-terminal portion of MLL in directing speckled nuclear localisation of the wild type protein, as well as of several aminoterminal truncations and chimeric fusions. Some of the molecular details responsible for this pattern were initially defined through overexpression assays, and therefore warrant confirmation in more physiological systems. This appears particularly relevant, since overexpression of either MLL truncations or MLL fusions has resulted in a somewhat different pattern, with bigger clusters as well as small dots and nucleolar staining. These data also bring a word of caution when designing experiments to assess the function of MLL and/or its fusion proteins under non physiological conditions, as such staining artifacts could well have a confounding functional counterpart.

IV.7 MLL function in the mouse

Two different *Mll* knock-out alleles have been generated.

In the first study (Yu et al., 1995), the function of *Mll* was ablated by fusing the LacZ reporter gene in frame with exon 3b, allowing simultaneous analysis of its expression pattern. In homozygous embryos this knock out design resulted in the absence of the full transcript, while a truncated transcript, coding for the N-terminus of the protein which includes the three AT hooks, was still detected.

Heterozygous mice displayed a complex phenotype, arguing that the *Mll* gene is haploinsufficient. However, as the translation of the truncated protein was not investigated, the possibility remains that at least some aspects of the phenotype could be due to a dominant negative effect. The phenotype was characterised by smaller size at birth accompanied by growth retardation, anemia and mild thrombocytopenia. B-cell populations were also reported to be consistently reduced. In addition, several segmental abnormalities were detected, which included: second cervical vertebral and sternal malformations, anterior transformations of cervical vertebra C7 to C6 and thoracic vertebra T3 to T2, as well as posterior transformations of T13 to lumbar vertebra L1 and L6 to sacral segment S1. Hence, one of the distinguishing characteristics of the *Mll*+/- phenotype, namely both anterior and posterior homeotic transformations, recapitulates the distinct homeotic phenotype of the first (hypomorphic) allele of trx identified in flies, namely bidirectional transformations of prothorax (T1) and metathorax (T3) towards a mesothorax (3 X T2) identity. RNA in situ hybridisation on +/- day 10.5 embryos showed that the expression boundaries of Hoxa-7 and Hoxc-9 were shifted caudally in both mesodermal and neuronal derivatives.

Homozygous mutants were embryonic lethal. *Mll*-/- embryos were still recovered at the expected Mendelian frequency between embryonic days (E) E9.0 and E12.5; however,

E10.5 was the latest timepoint for embryo viability, as judged by the presence of cardiac contractions. At this time, expression of *Hoxa7* and *Hoxc9* was completely missing in homozygous embryos. However, in these embryos, the correct spatiotemporal expression pattern of those genes was properly initiated, at around E7.5-E8.5, but failed to be maintained beyond E9.0. This strongly supports the view that, in remarkable analogy with fly Trx, MLL also maintains, rather than establishes, *Hox* gene expression patterns during embryonic development.

As the appropriate pattern of *Hox* gene expression defines the segmental identity of cell types along the anteroposterior axis, a detailed analysis was performed on cranial ganglia, spinal ganglia and somites (Yu et al., 1998). In good agreement with the maintenance model, neural patterning appeared to procede normally through the initial stages, but to then fail to establish the appropriate connections, especially as far as segmental patterning was concerned. As the segmental pattern of spinal ganglia is controlled by neighbouring somites, it came as no surprise that somites displayed disrupted architecture accompanied by widespread cell death.

In a further set of experiments, semiquantitative Reverse-Transcription (RT)-PCR was used to assess the expression of several *Hox* genes (from the *Hox*a and the *Hox*c cluster) in *Mll*-/- immortalised mouse embryonic fibroblasts (MEFs) from E10.5 embryos (Hanson et al., 1999). *Hox*a4 through *Hox*a10 were either absent or markedly reduced in *Mll*-/- MEFs, whereas *Hox*a3 was less affected and *Hox*a1 was normally expressed. Except for *Hox*c6, which showed reduced RNA expression, *Hox*c genes 4 through 9 were completely absent from *Mll*-/- MEFs. Caution should be taken in interpretation, since these cells were immortalised with the polyoma virus tumor antigen; however together these results strongly

support the idea that MLL maintains the expression of at least some genes of the *Hox* clusters.

The analogy with Trx function in flies is reinforced by the analysis of mice lacking both *Mll* and *Bmi-1*, the mammalian homologue of the *Drosophila* Polycomb group gene *Suppressor of Zeste 2 (Sz2)* (Hanson et al., 1999). Some of the skeletal abnormalities noted in the respective single mutant phenotypes were corrected when the two mutations were combined, most notably the widened C1 and the C2→C3 transformation accompanying Bmi-1 deficiency. However, the abnormalities of the sternum and the thoracolumbar spine were not corrected, arguing that the balanced loss of both proteins restores identity specification only in some segments. At the molecular level, *Hox*c8 and other, but not all, *Hox* genes were found to be reciprocally regulated by MLL and BMI-1, thus partially recapitulating the opposite actions of Trx and Pc in *Drosophila.*

A detailed analysis of yolk sac hematopoiesis was carried out on *Mll*-/- embryos (Hess et al., 1997). In this knock-out model, in fact, death preceded the full onset of liver hematopoiesis, which occurs at around E11.5-E12.5. Methylcellulose colony assays were performed on dissociated yolk sacs. Colony-forming-unit granulocyte, erythroid, macrophage, megakarycyte (CFU-GEMM), colony-forming unit macrophage (CFU-M), burst- forming unit-erythroid (BFU-E) and colony forming units-erythroid (CFU-E) were assayed. Except for CFU-E, which were normal or increased, in all other cases, a reduction in colony numbers, speed of growth and number of cells per colony was observed in *Mll*+/- embryos, which became more pronounced in *Mll*-/- embryos (60% decrease in the number of cells per colony). Importantly, mutant colonies did display the full range of differentiation seen in normal and heterozygous colonies, with a similar distribution of cell types, except for a striking reduction in monocytes and macrophages.

The second *Mll* knock-out experiment (Yagi et al., 1998) overall confirmed the hematopoietic disturbances described in the first study. In this case, *Mll*-/- embryos were reported to die between E11.5 and E14.5. Death was preceded and/or accompanied by edema and purpura. The delayed time of death, when compared to the first experiment, could be due to the adoption of different experimental conditions or to the difference in the targeting construct, which replaced exons 12 through 14 with a PGK-Neo cassette. These exons correspond to the breakpoint cluster region in human leukemias (which spans exons 8 through 14). A truncated transcript, encoded by the 5' end of the gene, was detected in *Mll*-/- embryos, which comprises five additional exons if compared to the mutant transcript generated in the first study, including the methyltransferase homology domain. Therefore, it was suggested that this longer truncated MLL version could result in delayed lethality. However, it should be noted that neither of the two studies investigated the presence of the truncated protein.

The delayed time of death enabled assessment of the initial stages of fetal liver hematopoiesis in *Mll*-/- embryos. Both histological analysis and colony forming assays from fetal livers were performed. The results largely recapitulate the situation observed with yolk-sac hematopoiesis: marked reduction in hematopoietic cells and their precursors with an overall normal differentiation profile along the various lineages. Again, granulocyte-macrophage colonies were particularly affected.

Postulating a defect in the proliferation and/or survival potential of *Mll*-/- hematopoietic precursors, an RT-PCR analysis was performed on several genes involved in cell cycle regulation, including *Rb*, *cyclinD1*, *cyclinD3*, *cdk2* and *cdk4*. No meaningful difference between wt and mutant embryos was detected. An analogous strategy revealed marked reduction in the expression of *Hoxa7*, *Hoxa10* and *Hoxc4*, in agreement with the previously reported results. On the whole, results on *Mll*-/- cells point to a pivotal function

for *Mll* in hematopoiesis; at the same time they also argue against a classical tumor-suppressor function for *Mll*, which would have predicted growth deregulation upon disruption of both *Mll* alleles.

IV.8 Mouse models of *Mll* leukemogenesis

IV.8.1 The *Mll-Af9* model

To model the *Mll-AF9* leukemia in the mouse, Rabbitts and coworkers used a knock-in approach to fuse the 3' terminal portion of the *Af9* cDNA to exon 11 in the endogenous *Mll* locus (Corral et al., 1996). Three other *Mll* mutant alleles were also generated: the AT-Lac allele in which the LacZ was fused in frame to *Mll* exon 3 (analogous to the knock-out construct described above), the *Mll*-myc-tag allele and the *Mll*-exon11-LacZ allel, in which a myc tag or LacZ, instead of the *Af9* cDNA, were fused in frame to *Mll* exon 11 (Corral et al., 1996; Dobson et al., 2000). Targeted ES cells for each mutant allele were injected into blastocysts. Chimeric mice, as well as heterozygous mice after germline transmission of the alleles, constituted the cohort in which leukemia development was assessed.

The great majority of both chimeric and heterozygous mice harbouring the *Mll-Af9* fusion allele developed acute leukemia, starting at 4 months of age, with very few mice surviving beyond 12 months of age. The total observation period amounted to 18 months (about half of the normal mouse life-span). Mice developed predominantly acute myeloid leukemias, but acute lymphoid leukemias were observed rarely , in agreement with the spectrum of *MLL-AF9* leukemias observed in humans. The disease bore a remarkable resemblance to the human condition, with bone marrow hypercellularity and extensive extramedullary hematopoiesis in spleen and liver. One of the most relevant findings was the pronounced expansion of the myeloid compartment observed prior to leukemia development in these

mice (Dobson et al., 1999). This was assessed by bone marrow fluorescence-assisted cell sorting (FACS) analysis at different ages using specific markers for the various hematopoietic lineages. Starting at birth, the *Mll-Af9* allele was associated with a clear proliferative advantage of myeloid cells (assessed with the myeloid markers gr-1 and Mac-1), which by 12 weeks of age amounted on average to 82% of the total bone marrow population in heterozygous animals versus 52% in wild type mice. This suggests that the *Mll-Af9* fusion ("the seed") is promoting the myeloid lineage ("the soil"), and that overt leukemia arises from an expanded myeloid pool as individual cells accumulate additional genetic lesions. This would also explain the relatively long latency of leukemia development in this mouse model, which had already suggested the need for secondary mutations in *Mll*-mediated tumorigenesis.

Importantly, the AT-Lac and the *Mll*-myc allele did not result in either benign myeloproliferation or overt leukemia, while mice harbouring the *Mll*-exon11-LacZ allele did develop acute leukemias. Interestingly, both the incidence and the latency of disease in this cohort were lower than for the *Mll-Af9* experiment (only 35% of the animals developed leukemia, starting at around 8 months of age, versus the *Mll-Af9* cohort where all the mice developed the disease). The distribution of the leukemic phenotype recapitulated the *Mll-Af9* results, with the great majority of animals suffering from acute myeloid leukemia and only three cases of lymphoblastic disease (either ALL or lymphoblastic lymphoma). Several insights can be drawn from these experiments.

First, it is possible to recapitulate in the mouse many aspects of *Mll* leukemogenesis, establishing this organism as a powerful platform to model other *Mll* translocations as well. The knock-in approach is in this regard a very useful tool. However, it still lacks features important for fully mimicking the human disease, the most important one being somatic acquisition of the translocation in the context of normal hematopoiesis. For example, the

extent of preleukemic myeloproliferation in *Mll-Af9* mice could well reflect the starting number of cells harbouring the fusion protein. These limitations emphasise the need to develop more accurate, fully conditional translocation models which rely on the Cre-loxP technology.

Second, mere truncation of *Mll* at either exon 3 or exon 11 does not seem to cause either leukemia or lineage expansion. On the contrary, in both AT-Lac (exon 3) and *Mll*-myc (exon 11) chimeras, mutant ES cells were preferentially excluded from the hematopoietic lineage. This did not correlate with the overall chimerism (as judged by coat colour), arguing that hematopoietic development is exquisitely sensitive to the correct dosage of the MLL protein. This conclusion is also in good agreement with the analysis of *Mll*-/- knock-out mice (Yagi et al., 1998; Yu et al., 1995).

Third, at least in the case of the *Mll-Af9* translocations, the fusion protein does indeed seem to direct lineage specification. In the light of these findings, the occurrence of lymphoid neoplasms with *MLL-AF9* fusions, both in these experimental models and in patients, can be interpreted as the result of additional genetic lesions which override the lineage-specific expression prompted by the *MLL-AF9* fusion. In addition, while the important role of MLL-fusion proteins in lineage determination is supported by animal models of the *Mll-Enl* translocation (see below), the *Mll*-exon11-LacZ results suggest that the myeloid lineage might be the default pathway for *Mll* tumorigenesis.

Possibly the most relevant observation from this set of experiments concerns the lack of leukemogenic potential displayed by the simple truncated *Mll* alleles (AT-Lac and *Mll*-myc) as opposed to the transforming properties of the *Mll*-exon11-lacZ allele. This has two immediate implications. The LacZ protein has been shown to assemble in a tetrameric complex (Jacobson et al., 1994), and it is therefore very likely that in this *Mll*-LacZ fusion it

provides an artificial homomerisation interface for the N-terminal half of MLL. Hence, in the spontaneously occurring diseases, homomerisation could be contributed by the different fusion proteins. At least two *MLL* translocation partners, *AF10* and *AF17*, contain leucine zipper dimerization domains, making this hypothesis at least plausible. A compatible possibility is that the LacZ primarily acts to stabilise the truncated MLL protein, a function which would normally be carried out by the C-terminal half of the protein. In either of these two scenarios, the MLL-LacZ fusion could then act in a dominant negative fashion and interfere with the normal MLL pathway. This hypothesis would partially incorporate the notion that MLL tumorigenesis simultaneously includes a loss of function (truncation) as well as a gain of function (domains contributed by the fusion partners).

At the same time, the observation that a LacZ fusion to exon 3 (which includes the three AT hooks) does not result in leukemia, argues that the methyltransferase homology domain and/or flanking regions are essential for the aberrant function of that fusion protein.

IV.8.2 The *Mll-Af19* model

These and other related questions were explored in the other major mouse model of Mll leukemogenesis. In this study, the oncogenic potential of the *Mll-fF19* translocation was investigated through transduction of hematopoietic progenitor cells with a retrovirus harbouring the *Mll-Af19* fusion cDNA (Lavau et al., 1997; Slany et al., 1998). As discussed previously, among all MLL translocation partners, *AF9* and *AF19* (also called *ENL*) share, together with *ELL*, some sequence similarity, and it is therefore particularly interesting to compare models for these two related fusion proteins. In these experiments, the effect of the fusion protein was monitored through methylcellulose colony forming assays. Infected cells showed an increased ability to generate myeloid colonies, which could undergo successive rounds of replating and finally established myelomonocytic cell lines. Upon transfer to both

sub-lethally irradiated syngeneic and severe combined immunodeficiency (SCID) mice, these immortalised cells gave rise to acute myeloid leukemia in the majority of animals. Similar results were obtained by transferring freshly isolated hematopoietic progenitor cells infected with the same virus. Mutant constructs were tested in a parallel set of experiments, to investigate which component of the -MLL-AF19 fusion was responsible for the in vitro immortalisation. The AT hooks and the methyltransferase homology domains, which are consistently present in all MLL fusion proteins, were both found to be necessary. Interestingly, the methyltransferase homology domain was shown to bind DNA through a south-western blotting assay using salmon sperm DNA and poly(dI-dC), which, like the AT hooks, may indicate a non-specific DNA binding property.

The AF19 component was also necessary for immortalisation (Slany et al., 1998). While expression of only the N-terminal half of MLL did not cause transformation, deletion analysis identified the 84 carboxy-terminal residues of AF19 as both necessary and sufficient for transformation when they were fused to MLL. This region has been predicted to contain two alpha-helices, and disruption of either one of them abrogated the oncogenic potential of the fusion protein. The region also corresponds to the segment of homology with Af9, and it is intriguing that in at least one documented case of *MLL-AF9* leukemia, only the carboxy-terminal 91 residues were retained in the fusion protein, which was therefore very similar to the minimal oncogenic fusion detected in these experiments. Therefore, *Mll-Af9* and *Mll-Af19* mediated leukemogenesis might procede along common pathways, which would be reflected in their preferential association with a myelomonocytic phenotype.

On the whole, these studies demonstrate the crucial role of a translocation partner for *Mll* leukemogenesis. They exclude the possibility that truncation may be the sole mechanism leading to tumorigenesis. While caution should be taken, since the deletion was analysed only by *in vitro* methylcellulose colony forming assays, the results demonstrating the role of

the AT hooks and the CXXC domain in immortalisation will help in the interpretation and design of further experiments.

Another interesting observation concerns the levels of expression of the fusion protein achieved in these studies. Only RT-PCR managed to detect expression of the fusion cDNA, while both northern and western blotting failed. Similarly to what has been reported for *PML*-RARα, the *Mll-Af19* cDNA also could not be expressed under the control of a constitutive promoter in a variety of cell lines. This indicates that only limited levels of the fusion protein can be tolerated in most cell types, and that this might turn out to be a more general feature of many translocation products. It also has far reaching implications for the design of translocation models, where alterations in the apparently low levels of expression required for transformation, as with standard transgenesis approaches, could dramatically affect the phenotype.

V

AF4

V.1 Introductory remarks

The *AF4* gene was originally identified through cloning of the translocation breakpoint in t(4;11)(q21;q23) cell lines (Corral et al., 1993; Djabali et al., 1992; Domer et al., 1993). At the time of its discovery, no other known genes showed any region of homology with *AF4*. Subsequently, three genes, which share high similarity with *AF4*, *LAF-4*, *FMR2* and *AF5q31*, were identified in mammals, (Gecz et al., 1996; Gu et al., 1996; Ma and Staudt, 1996; Taki et al., 1999), *AF5q31* being isolated because it is also translocated to *MLL*. They have been grouped in a new protein family of putative transcription regulators, termed ALF after the name of the most conserved domain shared among the members (from the initials of the three founder genes *AF4*, *LAF-4* and *FMR*-2). Very recently, a *Drosophila* homologue has been identified (Tang et al., 2001; Wittwer et al., 2001).

AF4 codes for a nuclear protein of 140KDa with a wide, possibly ubiquitous expression pattern (Chen et al., 1993). A diagram of the AF4 protein is shown in figure 2. It is particularly rich in serine and proline (26% of all residues), a feature interestingly shared by ENL, another of the *MLL* translocation partners. In addition, both proteins share a consensus GTP-binding site (GXXXXGK located at residues 946 to 952) (Morrissey et al., 1993; Morrissey et al., 1997). The main cDNA observed in most tissues is about 12.5 kb long, due to an unusually long 3' untranslated region. Alternative splicings have been detected in many tissues, as has the alternative use of two first exons (see below) (Morrissey

et al., 1993). This combinatorial potential produces proteins of different sizes (Li et al., 1998).

As for the other members of the family, their function is still largely unknown. *LAF4* was originally isolated from a Raji Burkitt's lymphoma cDNA library that was subtracted with K562 (erythroleukemia) cDNA to identify lymphoid specific genes (Ma and Staudt, 1996). It was found to be expressed in B cells, with the highest peak in pre-B cells and complete downregulation in plasma cells. It is a nuclear protein, with nonspecific DNA binding capacity, and two transactivation domains, one of which is conserved in AF4.

*FMR*2 is the gene associated with FRAXE mental retardation. FRAXE, located on Xq28, is one of at least five folate-sensitive fragile sites on the X chromosome, which are associated with various forms of X-linked mental retardation. Positional cloning of the FRAXE locus identified the causative mutation, for the majority of cases, as amplification of a CCG repeat immediately adjacent to the promoter of the *FMR*2 gene, in analogy to the mechanism described for FRAXA and *FMR*1. The expanded CCG repeat is methylated leading to silencing of *FMR*2 expression. An alternative mutational mechanism is found in some families that have been characterised, where microdeletions lead to severely truncated forms of the protein. Depending on the degree of CCG amplification, a range of premutation or full mutation alleles have been described, where full mutation correlates with the presence of mild non specific mental retardation (MR).

V.2 Protein domains present in AF4

Functional domains have been postulated in AF4, mostly on the basis of homology between the members of the family. A diagram of the AF4 protein and its domains is presented in figure 2. Expectedly, they are also the regions of highest homology between the mouse and the human protein (which have an overall homology of 64%). Two studies have reported such alignments, with largely overlapping results (Nilson et al., 1997) (Isnard et al., 1998). The differences, likely arising from the respective use of the human and mouse sequence of AF4, mainly concern the portion of the protein which contains the putative transactivation domain (Prasad et al., 1995). Whereas in the first report, aligning only the human sequences (Nilson et al., 1997), a single homology domain was proposed for this region (the ALF domain, see below), in the second study, which compared the mouse AF4 with the human AF4, FMR2 and LAF4 proteins (Isnard et al., 1998), two seemingly distinct domains of homology were identified, located respectively upstream and downstream of the BCR, with the downstream one including the transactivation domain. This last alignment is summarised here.

The first putative domain is located N-terminally (and is therefore called NHD), spanning 66 residues (position 6-72 in the mouse sequence) with 91% human/mouse homology. This region shares 53.8% homology with FMR-2 and 60% homology with LAF-4.

The second is located at positions 258-321 (mouse sequence), and displays 89% homology to the human counterpart. Homology to FMR-2 and LAF-4 is respectively 51% and 56%. These first two homology regions are located upstream of the translocation breakpoint, and are therefore not present in the MLL-AF4 fusion protein. The reciprocal fusion protein, which would harbour these two domains at the N terminus, has been detected

in most t(4;11)(q21;q23) leukemias, although its relevance to the leukemogenic process is completely unknown.

The third domain has attracted considerable interest, since it includes the putative transactivation domain identified through GAL4 fusion assays (Morrissey et al., 1997; Prasad et al., 1995). It comprises 105 residues (positions 405-510 in the mouse sequence) with 86% human/mouse conservation, including the serine/proline rich region. The degree of homology with FMR-2 and LAF-4 is 51% and 48% respectively. In LAF-4, it corresponds to one of the two strong transcriptional activation domains identified (Ma and Staudt, 1996). This domain is consistently retained in MLL-AF4 fusion proteins, and has therefore been postulated to aberrantly regulate putative MLL target genes. It is comprised in the ALF domain recognised by Nilson et al (1997).

The fourth region of homology is a short stretch of about 30 residues (positions 753 to 780 in the mouse sequence) with 86% similarity between human and mouse. Homology to FMR-2 and LAF-4 is 57% and 64%.

Finally, the fifth region of homology, located at the carboxyterminus (and hence called CHD for carboxyterminal homology domain), includes amino acids 968-1126, with 80% human/mouse homology and 48% and 52% similarity with FMR-2 and LAF-4 respectively.

In all other regions, homology between the different members of the family is below 40%. The homology between AF4 and AF5q31 is restricted to the first, third and fifth region, with homology values of respectively 62%, 57% and 48% (Taki et al., 1999). Of note, as with *AF4*, in the *MLL*-AF5 leukemia also the breakpoint is in the third region, upstream of the transactivation domain, and the phenotype is remarkably similar to the *MLL-AF4* leukemias.

Up to six putative nuclear localisation signals have been identified in the AF4 sequence, in good agreement with immunodetection of the protein in the nucleus (see below).

V.3 Organisation of the *AF4* genomic locus

All the information concerning the genomic organisation of this locus comes from the analysis of the human *AF4* gene, both in normal and in t(4;11)(q21;q23) leukemic cell lines. The gene is more than 100 kb long and contains at least 21 exons (Nilson et al., 1997) (figure 1). Two alternative exons have been described (exon1a and exon 1b), based on the isolation of distinct cDNAs differing at their 5' ends (Morrissey et al., 1997). For neither exon has the transcription start site been defined. The two exons are used respectively by the two major *AF4* transcripts identified to date, FelA and FelB (from the original name of the gene Fel). In addition, a third transcript has also been detected (termed FelC), which terminates about 1500 nucleotides into intron 3 due to a cryptic polyadenylation site (Nilson et al., 1997). Its presence has only been detected by RT-PCR, and the corresponding band was not visible by northern hybridisation, possibly indicating the very low efficiency with which this transcript is produced. Interestingly, though, similar truncated transcripts, coding for only the NHD and the ALF domain, have been reported for other members of the family (LAF-4Δ for LAF-4 and OX19 for *FMR*-2) (Chakrabarti et al., 1998; Ma and Staudt, 1996). The corresponding truncated proteins, which should however be cytoplasmic due to the lack of a nuclear localisation signal, remain to be identified before their functional relevance can be assessed.

The 5' part of the locus is characterised by the presence of large introns. Thus, exons 1a and 1b are at least 19 kb apart, and intron 1b is also larger than 10 kb. Intron 3 represents

the greatest portion of the breakpoint cluster region in t(4;11) leukemias, and has been estimated to be approximately 30 kb long. Less frequently, the breakage occurs in intron 4 or intron 5 (Nilson et al., 1997). All available data indicate that the genomic architecture of *AF4*, *FMR2* and *LAF4* is the same (Gecz et al., 1997). The similarity is particularly intriguing in the case of *FMR2*, which also features two very large introns, of about 150kb, at its 5' end (intron 1 and intron 3). The two patients from whom the gene was originally isolated actually had a submicroscopic deletion in intron 3, and two other cases of developmental delay are known, where the deletion was entirely within this intron (Gecz et al., 1996). Thus, it is intriguing to envision that this intron (and possibly the whole 5' region of the gene), both in *AF4* and in *FMR2*, contains unstable sequences which promote chromosomal aberrations, deletions in the case of *FMR2* and translocations in the case of *AF4*.

The significance of the *AF4* intron phase map (figure 1) for *MLL-AF4* translocations has been discussed in chapter IV.1.

V.4 Cell culture studies of AF4 function

The function of AF4 has been explored in cell lines by fusing different portions of the protein to the DNA binding domain of the yeast coactivator GAL4, and assessing transcriptional regulation of a reporter gene (Morrissey et al., 1997; Prasad et al., 1995). This analysis identified a putative transactivation domain which maps respectively to residues 480-560 and 365-572 in the two studies, and which is consistently retained in MLL-AF4 fusion proteins, hence the suggestion that they may contribute to leukemogenesis through

aberrant regulation of potential MLL target genes. As with MLL, this putative transactivation function displays both promoter and cell line dependency.

V.5 Expression pattern of *AF4*

The expression pattern of *AF4* throughout development and in adult life has been studied by northern blot analysis, western blot hybridisation and RNA in situ hybridisation. Combining the results of several different studies, a picture emerges of widespread expression.

These studies have also shown the presence of multiple transcripts and/or protein isoforms, resulting from alternative splicing and alternative use of at least the two first exons.

By northern blot hybridisation, human *AF4* RNA was detected in heart, kidney, skeletal muscle, placenta, pancreas, and a large variety of both hematopoietic and nonhematopoietic cell lines. In all cases, two transcripts were detected, of 10.5 and 12 kb.

Two studies in the mouse confirmed the nearly ubiquitous transcription of the gene (Baskaran et al., 1997; Isnard et al., 1998). In the adult mouse *Af4* RNA was present in all tissues examined. Thymus, lymphnodes and kidney expressed particularly high levels. Weaker signals were detected in spleen, bone marrow, heart, muscle, lung and liver, and the transcript levels were still weaker in testis and brain. However, there appears to be some discrepancy between the two studies in the relative levels of expression in liver, lung and brain, possibly reflecting the different probes used. All organs showed a band of 9.8-10 kb, in agreement with the human *AF4* size. A shorter band (4.6 kb) was also detected with longer exposures in all organs, probably representing an alternatively spliced transcript that is less abundant or is present in only a minority of cells in that organ. In addition, the kidney displayed unique transcripts.

These results were supplemented by RNA in situ hybridisation using a 300 bp probe from near the 3' end of the open reading frame of *Af4*, in a region not homologous to either *LAF-4* or *FMR*-2 (Baskaran et al., 1997).

With this probe, *Af4* appeared to have a distinct and specific pattern of expression in situ. In the heart, myocardium was negative, while the endocardium tested positive. *Af4* transcripts were abundant in hepatocytes, and rarer in the endothelial lining of the hepatic sinusoids. In the kidney, despite the strong northern hybridisation signal, transcripts were only found in epithelial cells of the medullary tubules and in cortical arteries. Glomeruli, including endothelial cells, were negative. In the pancreas, the gene was expressed in acinar cells, whereas both ducts and the Langherans islets were negative. In the gastrointestinal tract, both the smooth muscle and the lymphocytes of the lamina propria were positive, while epithelial cells were negative. Finally, the spleen, rich in extramedullary hematopoiesis, showed abundant *Af4* RNA in the myeloid, erythroid and megakaryocytic compartments within the red pulp. The white pulp, where mature B lymphocytes accumulate, had less RNA. Consistent with this, transcriptional downregulation was observed in the B-cell compartment as cells proceeded from the mantle zone to the marginal zone in their maturation pathway. This observation has an interesting counterpart in the analysis of *Af4* transcription during thymic development. Starting from E15.5, the expression of *Af4* in the developing thymus was assessed by northern blot analysis. On day E16.5, additional transcripts appeared, a longer one (12.kb), presumably corresponding to the one observed in other studies, and a 2.5 kb transcript, which has also been detected in some human cell lines. Curiously, the longer transcript was no longer detectable in the thymus of adult 6 weeks old mice. Cell lines in culture which corresponded to various stages of thymocyte development (from CD4-/CD8- to CD4+/CD8+ thymphocytes) were also positive for *Af4* RNA.

In development, a general pattern emerged of transcriptional downregulation of *Af4* during early embryonic differentiation. Embryos were analysed from E7.5 to E17.5. At E7.5, *Af4* RNA is detected across all germ layers, with lower levels in the neurectoderm and ectoplacental cone, and high levels in the mesoderm, visceral and parietal endoderm. The invading trophoblast is also positive at this stage. RNA levels increase up to day E9.5. Of note, *Af4* RNA is present in the yolk sac, which contains hematopoietic and endothelial cells at this time, as well as in the paraaortic splanchnic mesoderm, from where intraembryonic hematopoietic precursors are thought to originate.

At around day E10.5, transcriptional downregulation starts in some compartments, and is followed by further downregulation in other tissues as well, with the exception of the central nervous system, where RNA expression is widespread and sustained throughout development.

In almost every organ analysed, different cell types thereafter display different patterns of RNA expression, suggesting elements of tissue specific regulation. Thus, cardiomyocytes strongly express *Af4* RNA until day E11.5, but then stop transcribing it, while the endothelial lining of both the heart and great vessels continue to present *Af4* RNA. In the skin, *Af4* RNA is first expressed throughout all layers, then becomes restricted to the basal layer. In the gastrointestinal tract, both epithelia, lamina propria and smooth muscle express *Af4* initially, but eventually only the smooth muscle retains expression. Expression in skeletal muscle remains strong throughout development. In the lung, there is progressive transcriptional downregulation in the bronchiolar epithelium, and the transcript eventually disappears in the epithelia of the airways as well as in the endothelium of pulmonary vessels. Cartilage development follows a similar pattern, with high levels of expression in the perichondral cells and the chondrocytes of the primordial vertebral bodies, followed by marked downregulation in mature chondrocytes.

V.6 *Af4* function in the mouse

The function of *Af4 in vivo* was recently explored through a knock-out approach in the mouse (Isnard et al., 2000). The targeting vector was designed to delete the 3' end of exon 11 and the 5' end of intron 11. Homozygous mutant mice were born at the expected Mendelian frequency, indicating that the deletion was compatible with embryonal survival.

On the whole the phenotype observed was very mild and displayed low penetrance. Among homozygous mice on a mixed 129/BALB/c background, 20% were significantly smaller than wild type and heterozygous controls. The size retardation abated by 6 weeks of age, except when it was particularly severe and eventually led to death of the animal. The growth retarded mice also showed a twofold reduction in thymic cellularity. The spleen and the bone marrow were less affected. Flow cytometry analysis demonstrated that the deletion most likely affected immature CD4+ CD8+ double positive thymocytes undergoing positive selection, which were reduced by more than twofold. The levels of CD4 expression were halved in a fraction of both single (CD4+/CD8- or CD4-/CD8+) and double positive (CD4+/CD8+) cells. *In vivo* reconstitution of the single and double positive thymocyte compartments was assessed after glucocorticoid treatment, which selectively eliminates double positive cells by apoptosis. In the affected animals, thymocyte precursors (double negative) managed to repopulate the mature single positive cell compartment, though more slowly than wt and only reaching half the normal number of cells.

Analysis of the bone marrow revealed a threefold reduction in the size of the B-cell compartment, which mainly affected both preB cells (B220+, CD43- or B220+ IgMlow) and the more mature bone marrow B cells (B220+, IgMhigh). This suggests that the pre-B-cell receptor dependent expansion phase is impaired by this mutation, without a concomitant block of differentiation.

The effects of this *Af4* mutation support a potential role in both T and B cell development. The low penetrance of the phenotype could be due to the targeting strategy, as partially functional proteins could be produced by alternative splicing which skips exon 11. Hence, this exon 11 deletion allele is probably a hypomorph rather than a true knock-out. Alternatively, if protein function was completely abolished by the exon 11 deletion, the low penetrance indicates substantial redundancy in the physiological pathways in which *Af4* is involved. Combinatorial ablations of the *Af4* family members as well, or analysis of an unquestionable knock-out, will clarify these issues.

V.7 An *AF4* related gene in *Drosophila*

Very recently, three studies identified a gene in *Drosophila* which shares some regions of homology with the *AF4* family members (Tang et al., 2001; Wittwer et al., 2001). The gene, called Lilliputian (Lilli), is implicated in the control of cell size and also in the establishment of proper segmentation in the *Drosophila* embryo.

The predicted open reading frame of Lilliputian codes for a protein of 1673 amino acids. At the carboxyterminus, it features the CHD domain shared with members of the *AF4* family. The degree of identity when compared with different ALF family members varies between 31 and 37% over a region of 250 residues. Interestingly, the intron/exon boundaries of Lilliput in this region are conserved with genes of the *AF4* family, pointing to a common evolutionary origin for these exons.

Like *AF4*, *FMR2*, *LAF4* and *AF5*q21, Lilli has serine and proline rich regions, and the serines are concentrated at the same relative position as in the other family members. Whereas Lilliput and *AF4* share serine stretch and the CHD domain, both the NHD and the ALF domain homologies are missing in Lilliputian.

In addition, Lilli has two nuclear localisation signals, an AT hook domain, and a sequence homologous to a portion of a putative transactivation domain from the POU class of transcription factors.

Interesting functions have emerged for this gene from mutational analysis. Embryos lacking maternal Lilli have cellularization defects and display a pair-rule segmentation phenotype, thus establishing Lilli as a maternally active pair-rule gene. This effect probably results from its role in regulating expression of a distinct subset of early patterning zygotic genes, which include fushi tarazu (ftz), huckebein (hkb) and serendipity alpha (sry α). In contrast to zygotic pair-rule genes (which are expressed in seven stripes in the embryo), Lilli is not expressed in a segmented fashion, and probably does not function through the major gap genes knirps (kni), kruppel (Kr) or giant (gt), since the striped pattern of the primary pair-rule genes even-skipped (eve), hairy (h) and runt (run) do not appear to be affected. Furthemore, Lilli mutations phenocopy the sryα phenotype, where disruption of the appropriate cytoskeleton rearrangements leads to a defect in cellularization of the embryo.

The other major conclusion emerging from these studies is the role of Lilli in cell size determination. In the photoreceptor cell system, Lilli mutations result in a cell autonomous decrease in cell size. Similarly, ablating Lilli function in the eye and in the head by the ey-Flp system resulted in flies with reduced eye and head sizes. Two pathways are known to regulate cell size in *Drosophila*, the PI3K/PKB and the Ras/MAPK pathways. The Lilli mutant phenotype is similar to that observed for many components of both these pathways, but shows some distinctive features. Most notably, while overexpression of any of the molecules of these two pathways increases cell size, overexpressing Lilli results in reduced cell size, just like loss of function mutations. Moreover, mutations in the PI3K pathway result in a concomitant reduction in growth rate and proliferation, which is not observed for Lilli, and the Lilli mediated decrease in cell size affects only adult cells. Another hint to the

complex interaction of Lilli with the PI3K pathway comes from the genetic interaction with mutations in PTEN, a homologue of a tumor suppressor gene known to act as a negative regulator of the PI3K/PKB pathway; whereas tissues mutant for PTEN display both hyperplastic and hypertrophic growth, concomitant loss of Lilli partially restores normal cell size. The observation that the two mutations do not completely eliminate each other, though, argues that Lilli is not a simple downstream effector of the PTEN pathway, but rather that these two molecules might interact in more complex ways to regulate cell size.

In parallel, Lilli also exerts a partially redundant function in the Ras/MAPK pathway, where it seems to regulate the efficiency of signal transduction downstream of Raf. Analysis of different mutations indicate that the C-terminal domain of Lilli is crucial for its function downstream of Raf , but is less critical for its effects on cell size, thus possibly uncoupling the two functions. Remarkably, one of the mutants responsible for this "Raf phenotype" has a tyrosine to alanine mutation at a position in which the tyrosine is invariably conserved among all members of the *AF4* family.

Further characterisation of these Lilli mutants also is likely to shed more light on the function of the domain shared with AF4. It is intriguing to speculate that two of the major pathways regulating cell size and growth could both involve the C-terminal domain of AF4, and thus be decisively affected by the MLL-AF4 fusion protein. The essential function of Lilli for increased activity of the Ras pathway is in this regard particularly thought-provoking. Mutations which constitutively activate Ras are frequent in lymphomas, but absent in leukemias with *MLL* rearrangements implying that the aberrant function of the MLL-AF4 protein might partially circumvent the need for constitutive Ras activation (Mahgoub et al., 1998). Alternatively, the absence of Ras mutations in *MLL* leukemias might simply be one more example for the known, and poorly understood absence, or infrequency of this type of point mutation in leukemias.

VI

Advanced genome engineering:

new approaches and techniques

VI.1 Introductory remarks

The completion of large scale sequencing projects is radically changing the conceptual paradigms of molecular biology and medicine. The challenge ahead is to make sense of this unprecedented amount of information, a task which is usually referred to as functional genomics. This can be broadly shaped into two main areas of development.

On the one side, it is becoming increasingly possible to design strategies where the full biological output of a given experimental variable can be assessed at the same time. DNA microarrays are the most prominent example, enabling the documentation of entire genomic (for example in the case of single nucleotide polymorphisms) or transcriptomic (in the case of cDNAs) profiles for a specific set of biological conditions. Protein microarrays promise to constitute the next leap forward in this direction.

On the other side, mutational analysis is a prerequisite to explore biological function. It can be phenotype or genotype-driven. In the first case, the phenotypic effect of random mutations has to be traced back to the affected locus. In the second, prior knowledge of a gene is used to introduce desired mutations in order to alter the phenotype. In both cases, the newly available sequence information presents new opportunities and challenges for more efficient methods to manipulate genomes.

Improvements in our mutational capabilities together with the global dimension of current analytical tools are expected to greatly increase our understanding of biological phenomena, both in health and disease.

In terms of genotype-driven mutational analysis in higher organisms, one of the main problems is the complexity of the genomes, both in terms of the size of many individual genes, and of the long range interactions which have been shown to play a crucial role in gene regulation. Indeed, in many cases it may be worth asking what exactly is a gene, when some of the regulatory regions which influence its activity may lie very far away and even overlap with other loci. Rather than the conventional string-like representation of letters, it may be more faithful to visualise such genes as ordered nets connecting a centrum (the coding sequence) to its multiple regulatory elements at their various positions.

A prerequisite for mutational analysis, and at the same time a cornerstone of molecular biology, is precise methods to engineer and propagate cloned DNA for further analysis and/or use. However, the size of many eukaryotic genes constituted a formidable challenge, addressed in part by the development of several cloning vectors of large capacity, such as bacterial artificial chromosomes (BACs) (Shizuya et al., 1992), P1 vectors (Sternberg, 1992) and P1 artificial chromosomes (PACs) (Ioannou et al., 1994). These can carry sufficient DNA to include, in the majority of cases, whole eukaryotic genes and even gene clusters with the full set of regulatory regions. Although this solves the problem of cloning and propagation in *E. coli*, it would be of little use without adequate methods to efficiently and precisely manipulate the large DNA inserts. Traditionally, precise assembly of DNA molecules relied on restriction enzymes and DNA ligases, which are still in the mainstream of recombinant DNA technology. However, these techniques are limited to molecules of considerably smaller size, due to the frequent occurrence of cleavage sites. Furthermore, the necessity to utilise preexisiting restriction sites substantially narrows the range of modifications which can be introduced into a given molecule.

Polymerase Chain reaction (PCR) has obviated some of these problems and has greatly increased cloning flexibility. However, the size limitation and the mutagenicity inherent to any PCR approach seriously limit its application, when large portions of DNA need to be reliably and precisely manipulated.

A valid alternative to engineer DNA is presented by homologous recombination *in vivo* in *E. coli*.

VI.2 DNA engineering via homologous recombination

In homologous recombination, exchange of DNA occurs through regions which share sequence homology to yield a new recombined DNA molecule. It is a precise, faithful and specific reaction, which make it an ideal tool for DNA modification. This "technological power" has been long recognised, and two pivotal applications have been the introduction of desired foreign sequences at predetermined sites in the yeast and the mouse genome (Orr-Weaver and Szostak, 1983; Shashikant et al., 1998; Thomas and Capecchi, 1986; Thomas and Capecchi, 1987). The striking difference in the length of homology required in these two organisms for useful frequencies of homologous recombination (50 nucleotides versus more than 1 kb) points to the diversity of the pathways involved.

In the last few years, exploitation of homologous recombination in *E. coli* has advanced so that now the modification of any DNA molecule is possible. For cloning vectors like BACs, P1s and PAC this has meant a revolution in the ability to manipulate their large inserts in a precise, fast and predictable manner.

In any system (even yeast), homologous recombination is intrinsically a relatively rare event, which therefore requires an appropriate selection strategy to identify successful recombinants. The incorporation of a selectable marker (most commonly a drug resistance

gene) in the targeted DNA, however, is sometimes problematic. In order to remove it, several approaches can be considered:

1) The selectable marker can be flanked by recognition sites for site specific recombinases (SSRTs), so that, upon expression of the relevant site specific recombinase either in *E. coli* or in the final host of the construct, the intervening marker can be deleted. This is a commonly used procedure, and recently, the repertoire of recombinases available (traditionally Cre and Flp) has been augmented by the addition of TnpI (see chapter X.2.5) A potential drawback of this approach is that it leaves behind one 34 bp SSRT, which might be a problem if for example the coding region of a gene is being targeted.

2) The selectable marker can be flanked by restriction sites, so that it can be conveniently excised from the recombined molecule with the appropriate restriction enzyme. This strategy also leaves behind an obligatory stretch of sequence (though normally shorter and more flexible than a SSRT). As it relies on restriction enzyme digestion, it can only be applied to relatively small molecules.

3) A two step homologous recombination strategy can be implemented. In the first round, both a selectable marker as well as a counterselectable marker are targeted to the desired position of a DNA molecule. Selection of the first-round recombinants relies on the selectable marker. In the second round, both markers are eliminated by homologous recombination with a new linear DNA molecule which only carries the appropriate segments of homology. Selection of the second-round recombinants relies on the counterselectable marker. This is the cleanest strategy, as it yields a seamless product modified only in the intended way. However, the efficiency of the second round is much lower, since any mutation which ablates function of the counterselectable marker will score as positive.

Two methods for recombinogenic engineering in *E. coli* will be described, one based on recA and the other on ET recombination.

VI.2.1 RecA mediated homologous recombination

In E.coli, the major homologous recombination pathway involves the RecA protein (a strand invasion protein) and the RecBCD complex, which is the major cellular exonuclease amongst other activities pertinent to recombination. Despite the preeminence of *E. coli* as both a model organism and the cloning host of choice, homologous recombination has proven to be a difficult process to harness for DNA engineering. This is partly due to the fact that RecBCD plays a major role in distinguishing self DNA from foreign through recognition of an 8 nucleotide sequence, called Chi (χ), from double strand breaks. DNA without chi sequences is destroyed at the amazing rate of 1000 nucleotides/second. Consequently, it is not possible to introduce foreign DNA in linear form. However, double strand breaks greatly promote the initiation, and hence efficiency, of homologous recombination. There have been three solutions to this conundrum.

First, linear DNA is not employed (Hamilton et al., 1989; Yang et al., 1997). Rather, for every DNA engineering exercise, a dedicated targeting plasmid must be assembled which contains: the recA protein (to complement the usual absence of recA in cloning hosts); extensive regions of homology to the target molecule (usually around 1 kb); a selectable as well as a counterselectable marker; and finally a temperature sensitive (ts) origin of replication. In the first round of homologous recombination, carried out at a non permissive temperature for the plasmid origin, this whole plasmid is integrated into the target DNA molecule through the regions of homology. Selection of the recombinants relies on the selectable marker. The result is a cointegrate, which can be resolved by shifting to the permissive temperature. Under these conditions, resolution can occur through either one of

the regions of homology, resulting respectively in either the original, unrecombined or the intended recombined target molecule. Selection for this second step relies on loss of the counterselectable marker. This method has been successfully applied to the modifications of both regular plasmids, large vectors (BACs) (Yang et al., 1997) as well as to the E.coli chromosome (Hamilton et al., 1989). However, it requires the presence of large regions of homology, and hence the construction of a dedicated plasmid in a difficult, temperature sensitive backbone for each engineering step. This is obviously a great disadvantage if one wants to subject the same molecule (for example a BAC) to multiple rounds of modifications.

Second, Smith and coworkers reasoned that inclusion of chi sites next to double strand breaks would permit the use of linear, foreign DNA (Jessen et al., 1998). This strategy works but not with the anticipated efficiencies and the published reports indicate difficulties. Furthermore, this method, like all recA approaches, needs homology regions greater than 500 bp to accomplish workable efficiencies.

Third, Messerle found that the sbcBC host, which is recA+, recBCD-, did permit use of linear, foreign DNA. Since the sbcBC mutation activates the recA-dependent RecF pathway, this approach also requires long homology regions (Messerle et al., 1997).

Overall, all recA approaches also appear to deliver low ratios of homologous to non-homologous recombination. This may be due to the other activities of recA which include DNA repair and non-homologous end joining.

VI.2.2 ET mediated recombination

The alternative to recA based methods is ET recombination (Zhang et al., 1998a). Homologous recombination is initiated by either one of two protein pairs: RecE/RecT from

the Rac phage (after which the system is named), and Redα/Redβ from the λ phage (Muyrers et al., 1999; Zhang et al., 1998a). These two pairs have been shown to be functionally equivalent; RecE and Redα are 5'-3' exonucleases, while RecT and Redβ are DNA annealing proteins (Kolodner et al., 1994; Muyrers et al., 2000b). They not only serve to circumvent RecA, but also RecBCD. Consequently, *E. coli* hosts that are recBCD- can be used. Alternatively, RecBCD can be inhibited by expression of the λ phage Redγ protein. In fact, an unexpected outcome of the work that started with the discovery of ET recombination, is that this mechanism is the purpose of λ's operon (αβγ). Since RecBCD is removed, use of linearised DNA is possible.

In the fundamental reaction, a linear DNA molecule carrying a selectable marker flanked by regions of sequence identity is integrated at the desired location into a circular DNA molecule. The most important difference with the RecA method is that ET recombination requires only very short stretches of homology (in the range of 40-60 nucleotides), which can be easily incorporated into oligonucleotides. Indeed, ET recombination appears similar to homologous recombination in S. cerevisiae in several ways, with perhaps the most useful being the high ratio of homologous to non-homologous recombination. In the simplest case, the recombinogenic, linear DNA molecule can be generated by a PCR reaction to amplify the desired selectable marker with oligonucleotides consisting of two parts: a 3' PCR primer and a 5' part consisting of the 40-60 nucleotides of sequence identity with the chosen site in the target molecule. No dedicated plasmid need be assembled, rendering the whole engineering exercise much more fluid.

ET recombination has been successfully applied to regular plasmids, large cloning vectors (like BACs) and the *E. coli* chromosome. It shows consistently high efficiencies in all these applications. Usually more than 80% of recombination events are homologous.

A remarkable development of this methodology has recently been its application to direct cloning and subcloning (Zhang et al., 2000). In this set up, the linear, recombinogenic DNA molecule is a PCR generated plasmid backbone, carrying a selectable marker and an appropriate origin of replication. As for any ET experiment, the PCR oligos include regions of sequence identity which precisely delimit the boundaries of the DNA molecule to be cloned or subcloned. Upon homologous recombination, the desired portion of the target DNA molecule is faithfully copied into the plasmid backbone. This method has been applied to subclone desired segments from regular plasmids, BACs and even the *E. coli* chromosome. It presents several interesting uses. First, in a variety of cloning exercises, it can conveniently transfer a desired DNA region from large cloning vectors to higher copy, easier to handle cloning vectors. Second, it can be combined with conventional cloning methodologies, by inserting appropriate restriction sites in the homology arms, which will now flank the selected segment for subsequent cloning steps. Specifically, the subcloning from a BAC of the whole third intron of the mouse *Af4* gene presented in this thesis pioneered the use of this technology and demonstrated that it can be used to precisely subclone very large DNA segments (28 kb in this case). This presents two chief adavantages over PCR (which has been used to achieve similar goals), namely the wider size range over which it can be applied, and the very high fidelity of the reaction. In fact, recombination occurs in vivo in *E. coli*, and is therefore subjected to the physiological proof-reading properties of *E. coli* DNA replication machinery, limiting drastically the occurrence of mutations.

Remarkably, the same principle has been successfully applied to the direct cloning of a desired DNA molecule from a complex mixture such as genomic DNA preparations from E.Coli, yeast and mouse. The frequencies at which the correct products were obtained in these three cases were respectively 100%, 33% and 17%, broadly relating to the increased

complexity of the starting population of DNA fragments. In this set up, ET recombination is conceptually analogous to PCR, in that it enables the amplification of a desired DNA region from a complex mixture. Optimizations in several reaction parameters are likely to further increase the efficiency, unfolding new possibilities to directly clone pieces of the genome without the need of library screening. The availability of complete genome sequences will soon eliminate the problem posed by the requirement for prior knowledge of the sequences flanking the target region.

In this thesis, three basic ET cloning variations are presented to assemble complex, versatile targeting constructs and BAC transgenes to be used in mouse functional genomics. They are schematised in figure 3.

Panel A shows the basic ET reaction in which a PCR product is inserted at a desired location in the target molecule. Importantly, this strategy works independent of the distance intervening between the homologous regions on the target molecule. This means that it can be used either to insert a foreign sequence between adjacent nucleotides, or, at the other extreme, to delete very large portions of intervening, undesired sequence; in this last case, the PCR product replaces the selected region resulting in a net loss of genetic material. This strategy was applied to the characterisation of the *Mll* BAC (see chapter VII) and to the assembly of the final BAC based *Mll* targeting construct.

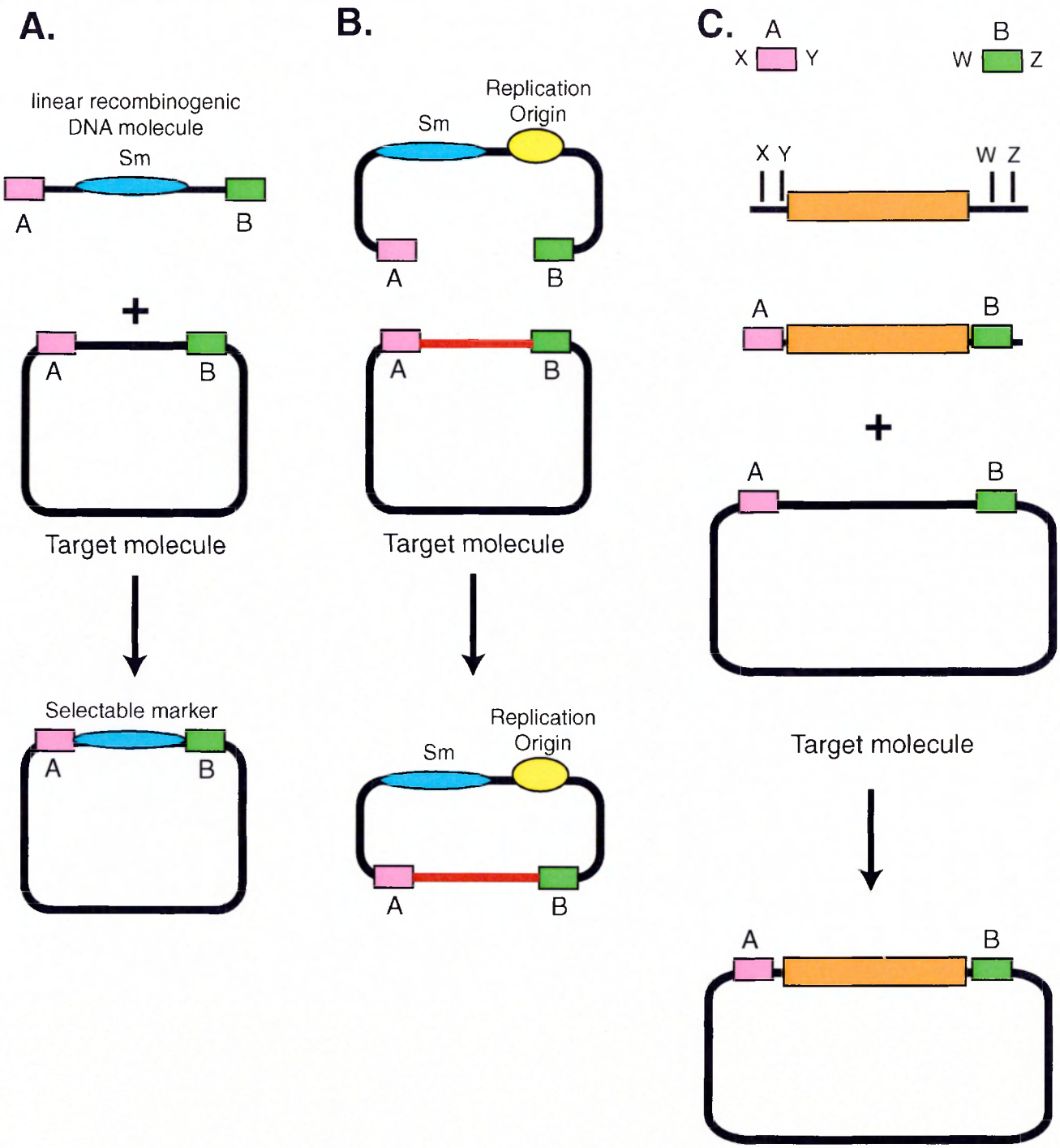


Figure 3 Schematic illustration of the three main variations of the ET recombination technology applied to the generation of the targeting constructs described in this thesis
Panel A illustrates the most basic setting, in which a PCR product carrying arms of homology is precisely integrated at the desired location in the target molecule.
Panel B depicts the use of the same methodology to precisely subclone a fragment of interest from a donor plasmid (or a BAC) into a more convenient acceptor vector. The acceptor vector is amplified by PCR including the appropriate arms of homology, in analogy to the strategy depicted in panel A.
Panel C illustrates a variation of the strategy of panel A, in which the homology arms are not incorporated via PCR, but rather cloned by conventional methods into a pre-existing targeting cassette, to yield the final linear DNA molecule used in the ET reaction. This strategy avoids the problem of mutations generated during the PCR reaction.

In panel B, the basic aspects of ET subcloning are depicted. This technique was instrumental to the characterisation of the mouse *Af4* BCR through subcloning intron 3 from a BAC, and subsequent subcloning from the BAC the chosen portion of the *Af4* gene for the *Af4* targeting construct.

Panel C shows a variation of the basic ET cloning strategy depicted in figure 3a, which circumvents the mutational risk inherent to PCR amplification (Angrand et al., 1999). This is an important consideration when, most typically in the assembly of mouse targeting constructs, several functional elements (selectable markers, protein tags, reporter genes, splice acceptor and IRES elements) need to be incorporated in the linear DNA molecule. While deleterious mutations for some of these (double selectable markers for use in both *E. coli* and eukaryotic systems) can be easily identified, for most the functional test would have to await the mammalian experiment, and this constitutes an unacceptable risk. Therefore, to avoid PCR, the desired targeting cassette is first assembled in a regular plasmid through conventional and/or ET cloning, and suitable restriction sites are placed at both of its ends. Homology arms for ET recombination are then cloned by conventional ligation at the two sides of the cassette, which thus does not undergo any round of PCR. Prior to the ET reaction, the recombinogenic DNA fragment is excised from the plasmid by restriction digest. This necessarily results in the presence of non homologous sequence at the ends, minimally the restriction sites. However, this does not interfere with the method, since, as expected, homologous recombination eliminates these spurious overhangs upon integration into the target loci. This strategy was applied to generate all selectable marker cassettes described in this thesis, which were then targeted by ET recombination to the three BACs encoding respectively the *Af4*, *Mll* and *Ikaros* gene.

RESULTS AND DISCUSSION

VII
The objective of this thesis

The objective of this thesis was to develop and apply novel genome engineering strategies for the assembly of complex targeting vectors to model in the mouse the acute leukemia associated with the t(4;11)(q21;23) translocation. The chosen approach based on the Cre-loxP technology (Figure 4) requires that the two genes have the same centromere-telomere orientation, otherwise, upon Cre recombination, an acentric and a dicentric chromosome would be generated and impair cellular viability. The still very limited amount of sequence information available for the mouse genome (1.3% of the whole genome as finished sequence and 10.3% as draft sequence) does not allow yet to determine the orientation of the *Mll* and *Af4* genes in the mouse with bioinformatics tools. I relied on the following genetic mapping data to conclude that the likelihood that the two genes had the same orientation was sound enough to warrant a Cre-loxP based approach.

The *Mll* gene was mapped to mouse chromosome 9 by interspecific backcross analysis (Ma et al., 1993) at 26.0 centimorgan (cM) from the centromere, tightly linked to the Cd3d locus (< 2.9 cM apart). This was in agreement with the human location, as a probe for the human homologue of Cd3 was originally used to identify by positional cloning the human *MLL* gene (Cimino et al., 1991; Cimino et al., 1992; Ziemin-van der Poel et al., 1991). The *Af4* gene was mapped to the central region of mouse chromosome 5 by interspecific backcross analysis, at 56 cM from the centromere in a region of conserved sinteny with the human 4q21 region (Baskaran et al., 1997). It is linked to the Alb1, fgf5 and Adrbk2 loci, and based on the backcross results, the most likely gene order is: centromere-

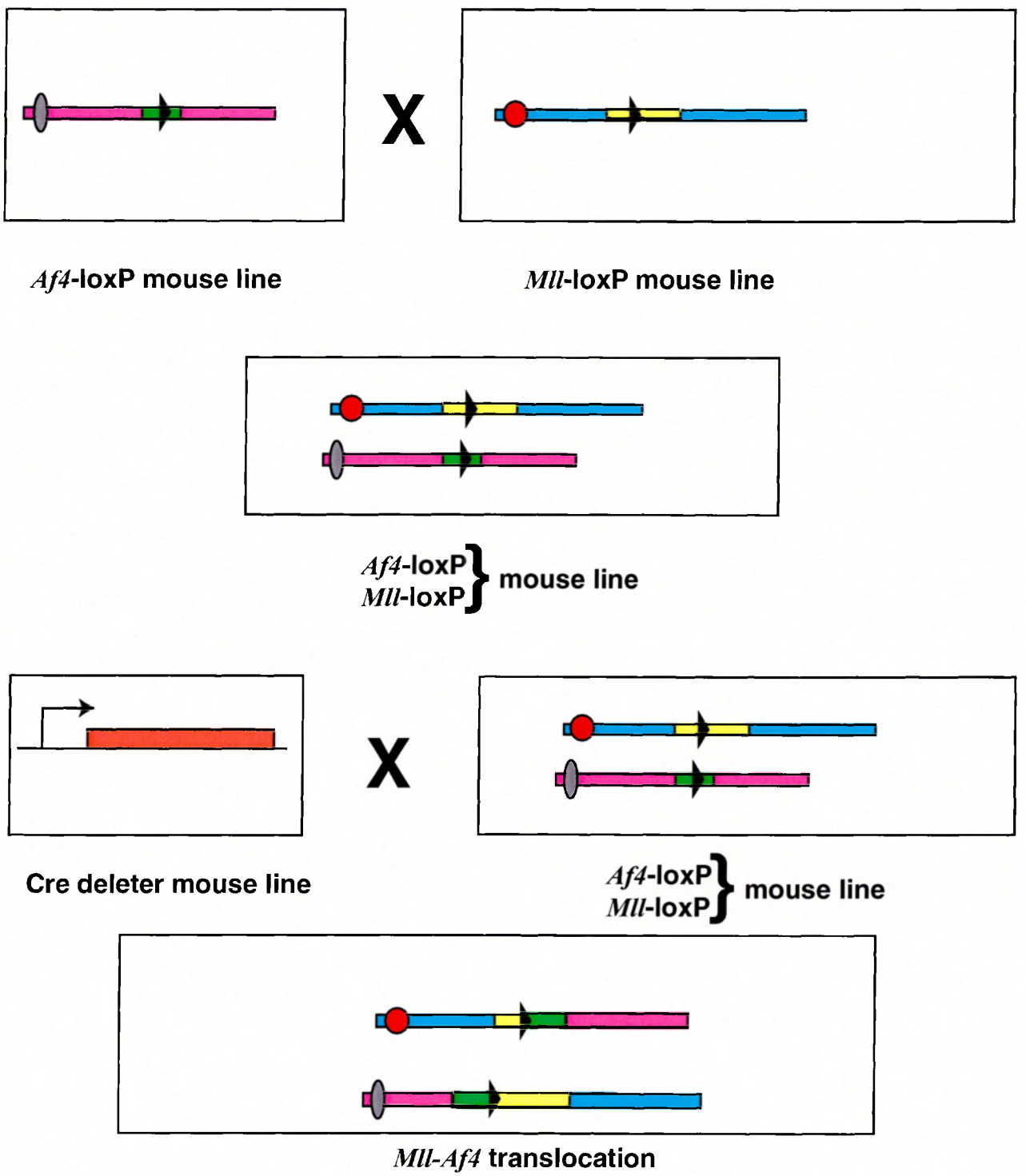
Alb1-Fgf5-*Af4*-Adrbk2. Location of a gene in a conserved region of sinteny between genomes is in itself not a guarantee that the genes have maintained the same order. Comparison of the human genome with the genome of the pufferfish *Fugu rubripes* has in fact revealed that whereas some regions of conserved sinteny display a striking conservation of gene order(Miles et al., 1998), others do not(Gilley and Fried, 1999). However, the overall evidence from the backcross analyses reported argues that both genes have the same centromere-telomere orientation. Importantly, this assumption has been strengthened in the case of *Mll* by the study reporting the Cre induced *Mll-Af9* translocation (Collins et al., 2000), which constituted an indirect evidence that the two genes, both of which had been mapped to regions of conserved mouse/human sinteny, had maintained the same orientation in the mouse.

Complex alleles were then generated for the *Mll* and *Af4* genes, which allow them to be used for Cre-driven interchromosomal translocation, as well as to study the function of these genes in the mouse. In order to drive the *Mll-Af4* interchromosomal translocation, a transgenic mouse line was established in which Cre is expressed under the control of the *Ikaros* gene. The strategy employed should allow both regulated as well as constitutive Cre activity.

The following sections describe the results obtained for each line (*Af4*, *Mll* and *Ikaros*-Cre), and discuss the experimental strategies adopted and their perspectives for advanced genome engineering.

Figure 4 Diagram of the Cre-loxP approach to model the t(4;11)(q21;q23) leukemia in the mouse.

At least three mouse lines need to be established, two harbouring a loxP site in the relevant intron of the *Af4* and *Mll* genes respectively, and the third driving the expression of Cre recombinase. Various Cre-expressing mouse lines can be used, to assess the time and/or tissue dependence of the oncogenic activity of the fusion protein. LoxP sites are indicated by black arrowheads. The *Af4* and *Mll* genes are depicted respectively in green and yellow.



VIII
Engineering of a multifunctional *Af4* mouse line

VIII.1 Overview of the strategy

The *Af4* targeting construct was engineered from a bacterial artificial chromosome (BAC). BACs offer many advantages for performing complex DNA engineering exercises, the two main ones being the large size of the cloned insert (up to 150kb) and the convenience of handling BACs compared with other large cloning vectors like yeast artificial chromosomes (YACs). In the case of the *Af4* gene, it was known that human *AF4* is a long gene (about 100kb), and that the breakpoint cluster region involved in the t(4;11)(q21;q23) translocation has been mapped in the majority of cases to the third intron, which in humans has an estimated size of 30 kb (Nilson et al., 1997). It was therefore desirable to obtain as large a clone of the murine homolog as possible, which would span at least the whole of intron 3, in order to characterise the appropriate region for the mouse targeting experiment.

In assembling the targeting construct for *Af4*, I aimed at a versatile, multifunctional cassette, which could provide multiple entry points to analyse the function of this gene.

The primary goal was to introduce a loxP site into the third intron of the *Af4* gene for Cre-mediated interchromosomal translocation in the mouse. This single step could in principle be achieved by targeting the relevant region of this intron with a relatively simple cassette, featuring a selectable marker, most commonly the neomycin resistance gene, flanked by two loxP sites. Upon Cre recombination either in ES cells or directly in the mouse, the selectable marker would be deleted, leaving behind one single loxP site in the chosen site of intron 3. However, this targeting strategy would only enable the study of *Af4* function as an *Mll* partner in *MLL-AF4* associated leukemias. In spite of the fact that the

MLL-AF4 translocation is the most frequent *MLL* associated translocation, very little is known about the function of *AF4* per se. The homology between *AF4*, *LAF-4* and *FMR2* (see chapter IV) has identified a potentially new family of genes, but other than the conserved domains shared by all three members, no regions of similarity to other known proteins have been found. The observation that *MLL-Af4* leukemias almost invariably have an early lymphoid, most often Pre-B phenotype suggests a function for *Af4* in lymphopoiesis. According to the prevailing hypothesis the AF4 moieties present in the MLL-Af4 fusion protein direct a cell in which the translocation has occurred towards lymphoid differentiation.

Thus, it appeared useful to combine the requirements of placing a loxP site in the third intron of the gene with a multi-purpose knock-out strategy. Towards this end, the targeting cassette included a fusion of the lacZ to the neomycin resistance gene (βGeok), behind a splice-acceptor site and an internal ribosomal entry site (IRES) element, followed by the strong SV40 late polyadenylation signal. The βGeok cassette used was based on the βGeo construct published by Smith and coworkers (Mountford et al., 1994). βGeo comprised intronic sequences and a splice acceptor site from engrailed 2 (en-2) and the encephalomyocarditis virus (EMCV) internal ribosomal entry site (IRES) fused to a lacZ-neomycin resistance fusion gene. Previous work in the Stewart lab altered the βGeo cassette to create βGeok, by (i) changing the 50 aa. linker region between the end of lacZ and start of neo to include, whilst keeping the open reading frame, an *E. coli* promoter and ribosome binding site (rbs). The purpose of this change was to permit expression of the neomycin resistance gene in *E. coli* so that the βGeok cassette would deliver kanamycin resistance. Consequently, integration of the cassette by ET recombination could be selected by kanamycin resistance in *E. coli* and then used for G418 selection in ES cells. (ii) The polyadenylation signal of βGeo was replaced with that of SV40 late. The SV40 late

polyadenylation region, comprising more than 200 bp, is probably the strongest polyadenylation element described. Earlier work in the Stewart lab confirmed that it is the most reliable polyadenylation element for complete processing of the nascent transcript in many contexts. Hence, SV40 late polyadenylation can be inserted at the 5' end of a gene to create a knock-out allele through RNA processing without the need to delete segments of the gene in question. In different experimental settings, the βGeok cassette has proven efficient in yielding ES cell and mouse lines. βGeok simultaneously achieves, in one targeting event, monitoring of the expression pattern of the gene being studied through βgalactosidase activity and a gene knock-out. For the *Af4* translocation allele, the βGeok cassette was flanked by loxP sites. A further modification of the loxP-βGeok-loxP cassette was made for the following reason. The βGeok cassette can only be used for homologous recombination in active genes in ES cells because selection for neomycin resistance relies on an IRES. Consequently, I examined *Af4* expression in ES by RT-PCR and Northerns. Although *Af4* mRNA was detected in ES cells by RT-PCR, Northern analysis failed to detect any ES cell *Af4* transcript. Therefore, a hygromycin resistance gene, expressed from the SV40 promoter was included between the SV40 late polyadenylation region and the 3' loxP site, so that targeting into the potentially inactive *Af4* gene in ES cells could be selected by hygromycin resistance. Ironically, I subsequently found that this precaution was unnecessary, since the *Af4* gene in ES cells expresses sufficient mRNA for neomycin selection from the homologously integrated βGeok cassette.

After subcloning and sequence characterisation of the *Af4* BCR (see next section), a target site for homologous recombination in intron 3 was chosen. Since the loxP-βGeok-SV40hygro-loxP cassette was too large to be reliably amplified by PCR, the short homology arms for ET recombination with an *Af4* intron 3 subclone were added by conventional

ligation of *Af4* intron 3 PCR products at either side of the loxP sites (strategy depicted in figure 3, panel c).

A mouse line carrying the loxP-βGeok-SV40hygro-loxP cassette inserted into the *Af4* gene will incorporate four functions (figure 5)

1) Prior to Cre recombination, the sA-IRES-βGeok-pA cassette permits convenient display of the transcription of the gene by staining for β-galactosidase activity.

2) Prior to Cre recombination, the sA-IRES-βGeok-pA cassette can result in partial or total ablation of protein function through complete polyadenylation of the nascent transcript.

3) Following Cre recombination, the *Af4* allele will be wt but the loxP in intron 3 can be used for interchromosomal translocation with the *Mll*-loxP allele.

4) The loxP-flanked sA-IRES-βGeok-pA cassette is amenable to lineage specific gene repair strategies using mouse lines expressing Cre recombinase in a tissue specific fashion.

The experimental steps involved in generating this targeting construct are summarised below:

1. Isolation of a mouse *Af4* BAC.

2. Characterisation of the relevant regions of the BAC through:

    2.1 Restriction mapping Southern hybridisation

    2.2 Direct BAC sequencing

3. Assembly of the *Af4* ES targeting construct by ET recombination:

    3.1 Assembly of the ET targeting cassette bw-sA-IRES-βGeok-pA-hygro-SV40-aw

        3.1.1. Generation of PUX4-sA-IRES-βGeok-pA-hygro-SV40, a universal knock-in targeting vector which is independent of the expression of the target gene in mouse ES cells.

3.1.2. Assembly of the final *Af4* ES targeting vector PUX4-bw-sA- IRES-βGeok-pA-hygro-SV40-aw

3.2 Subcloning of the targeting backbone from the *Af4* BAC (*Af4*sub)

3.3 Insertion of selectable marker cassette bw-sA-IRES-βGeok-pA-hygro-SV40-aw into the desired region of the targeting backbone from the *Af4* BAC (*Af4*sub).

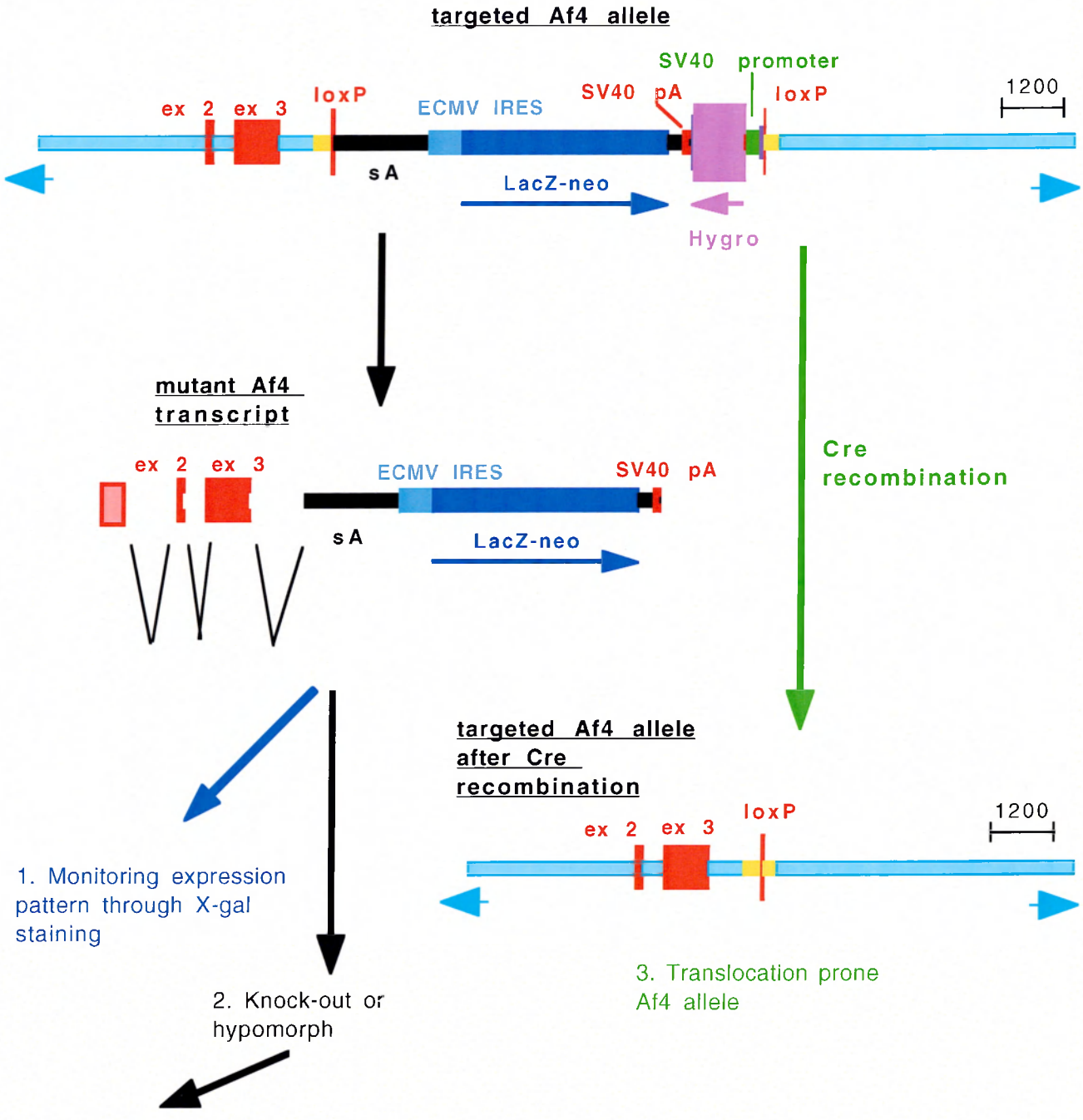


Figure 5 Potentials of the *Af4* recombinant allele
The targeting construct for the Af4 gene incorporates four functions. Prior to Cre recombination, it can report expression of the gene through X-gal staining, and can potentially result in a loss-0f-function phenotype, which would be amenable to conditional gene repair strategies. Following Cre recombination, it results in a translocation prone allele. Azure arrows indicate flanking genomic sequences.

VIII.2 Assembly of the *Af4* targeting construct

VIII.2.1 Isolation of the mouse *Af4* genomic clone from a high density BAC library

High density mouse BAC filters (Research Genetics, Inc.) were screened with a probe specific for the mouse *Af4* gene. The probe was a 600 bp fragment of exon 3 derived by PCR from mouse genomic DNA. The sequence of exon 3 for primer design was derived from the Genebank entry of mouse *Af4*. Primer *Af4*ex3F has the following 5'-3' sequence:

CTTTATCGATTGGGTGACTATGAGGAGATGAA.

Primer *Af4*ex3R has the following 5'-3' sequence:

TATCTCTAGAAGGGGAGAGGAGGGGGTGGAAA.

Hybridisation of this probe (*Af4*ex3) to the high density BAC filters yielded two independent positive signals (Figure 6). Both clones, n.189o7 and 183c14, were screened by PCR and Southern hybridisation and found to contain the third exon of the mouse *Af4* gene (Figure 7) The correct identity of both *Af4* BAC clones was further determined by directly sequencing the two BACs with a primer annealing to exon 3 and reading in the direction of intron 3. After confirmation of the correct identity of the two BACs, one BAC (n.189o4) was chosen for further characterisation.

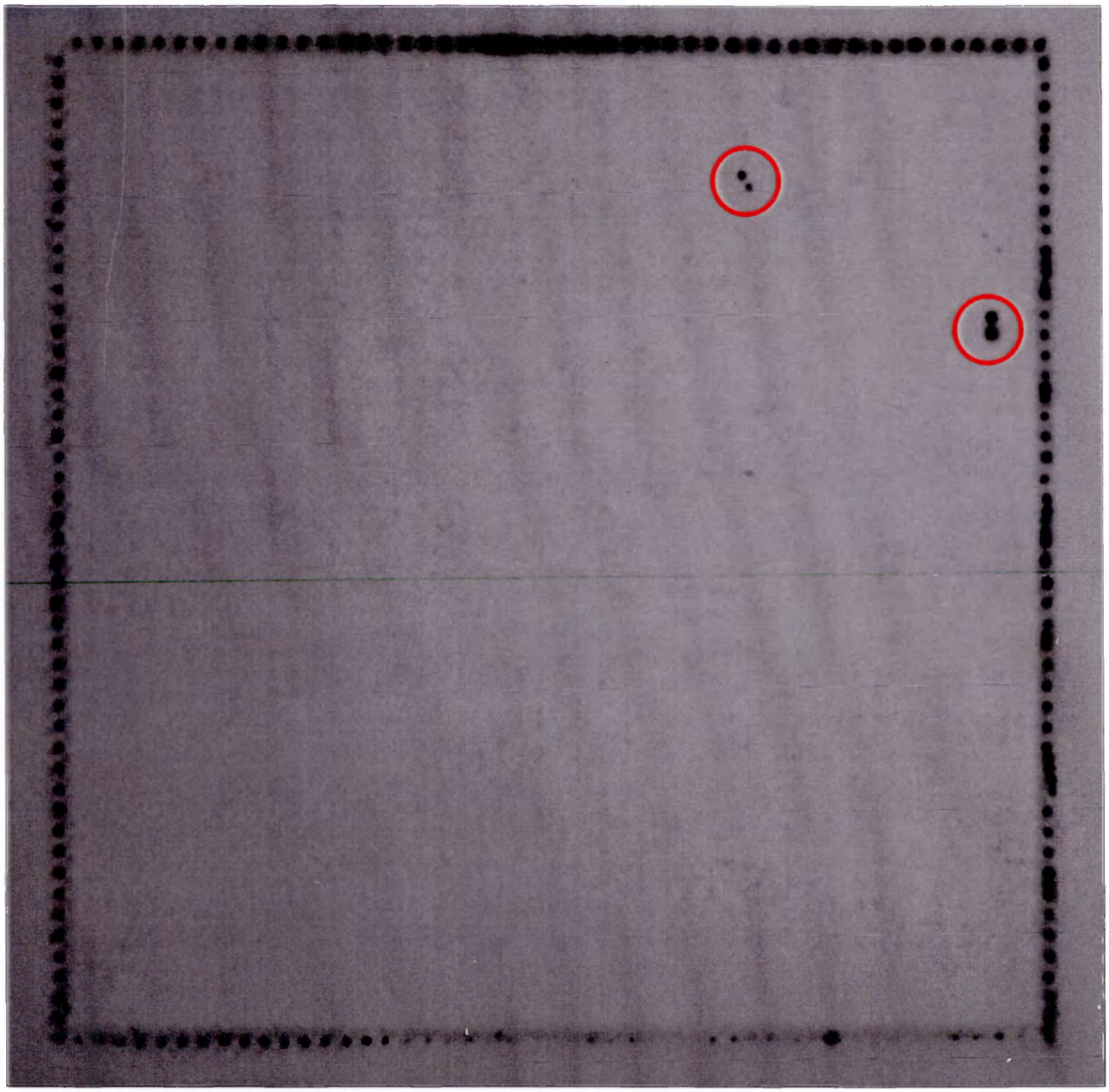


Figure 6 Isolation of the Af4 BAC from a high density BAC library
Hybridisation of high density mouse BAC filters with an Af4 probe (ex3) spanning approximately 600 bp in exon 3. Each BAC clone is spotted twice to distinguish true signals (a double spot) from possible false positives. The filter was simultaneously hybridised with a control probe which highlights the reference marks on the four sides of the membrane to unequivocally identify the clone of interest.
Positive signals are circled in red.

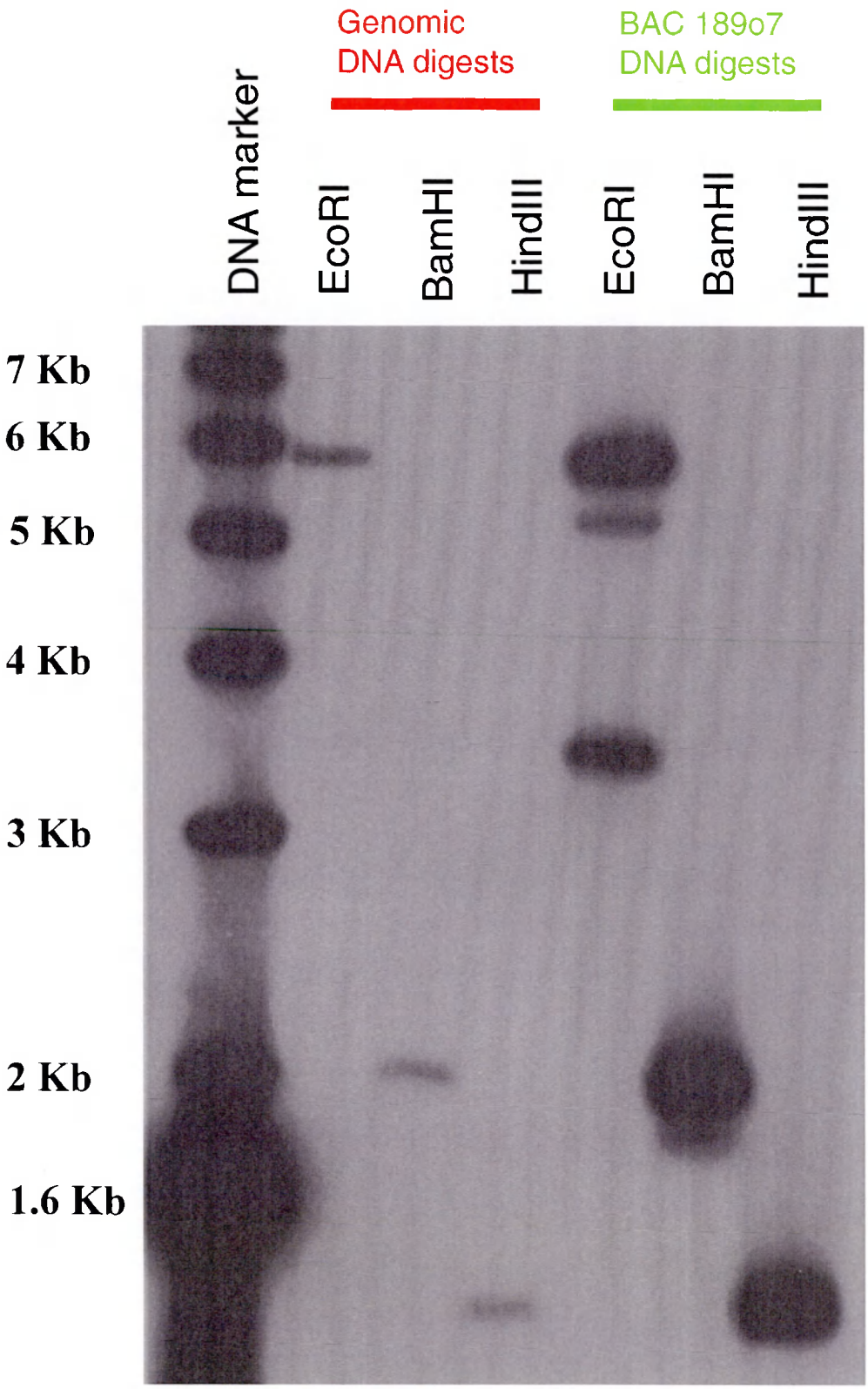


Figure 7 Mouse Af4 genomic and BAC blot
Southern blot hybridisation of genomic DNA and BAC DNA with the Af4 exon 3 probe to verify the correct identity of the BAC 189o7 isolated from the screening of the mouse BAC library.

VIII.2.2 Restriction mapping and Southern hybridisation

Two probes were used to generate a restriction map of the region of the *Af4* BAC spanning intron1 through intron3. The first probe was the same *Af4*ex3 probe used to isolate the BAC clone (see above). The second probe (g51) spanned 624 residues of intron 3, 2349 bp from the end of exon 3. The sequence information necessary to amplify this probe by PCR was derived from the sequencing strategy outlined in the next section.

DNA from BAC 189o7 was digested with a battery of restriction enzymes. Single restriction results were complemented by double restriction digests with NheI as a reference enzyme. Figure 8 shows the Southern hybridisation results from a representative set of digests. These data were then assembled in a restriction map of this region of the BAC (figure 9)

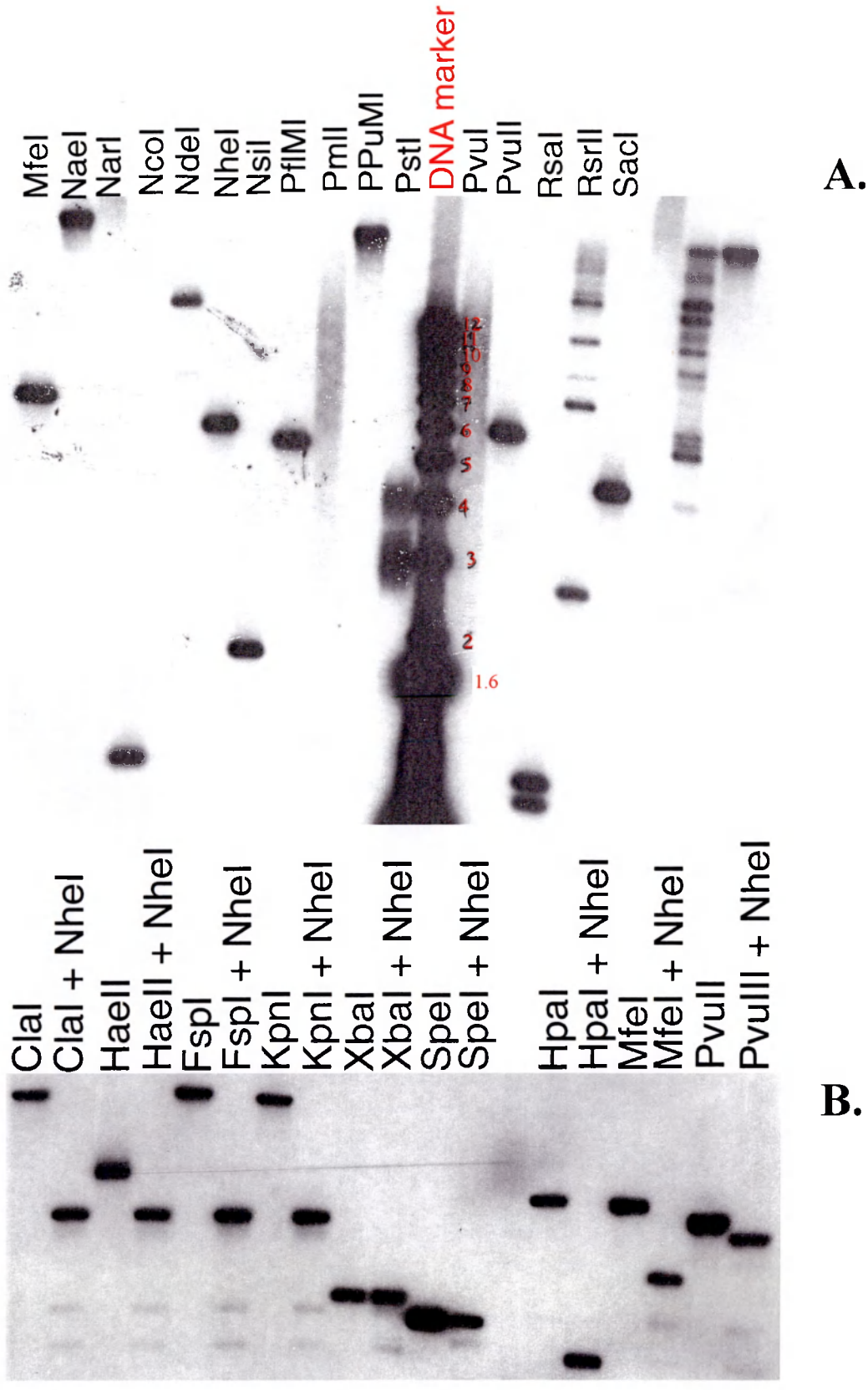


Figure 8 Southern hybridisation restriction mapping of the Af4 BAC
Representative single (**A**) and double (**B**) digests of the Af4 BAC 189o7 used to map the region corresponding to the breakpoint cluster region (BCR) in human leukemias. This was instrumental to develop an appropriate strategy to target the loxP flanked βGeok cassette, and to screen for homologous recombinants in mouse ES cells. The figure represents two blots hybridised with the ex3 probe, which spans about 600 bp in the third exon of the Af4 gene.
The size of the marker bands is expressed in Kb.

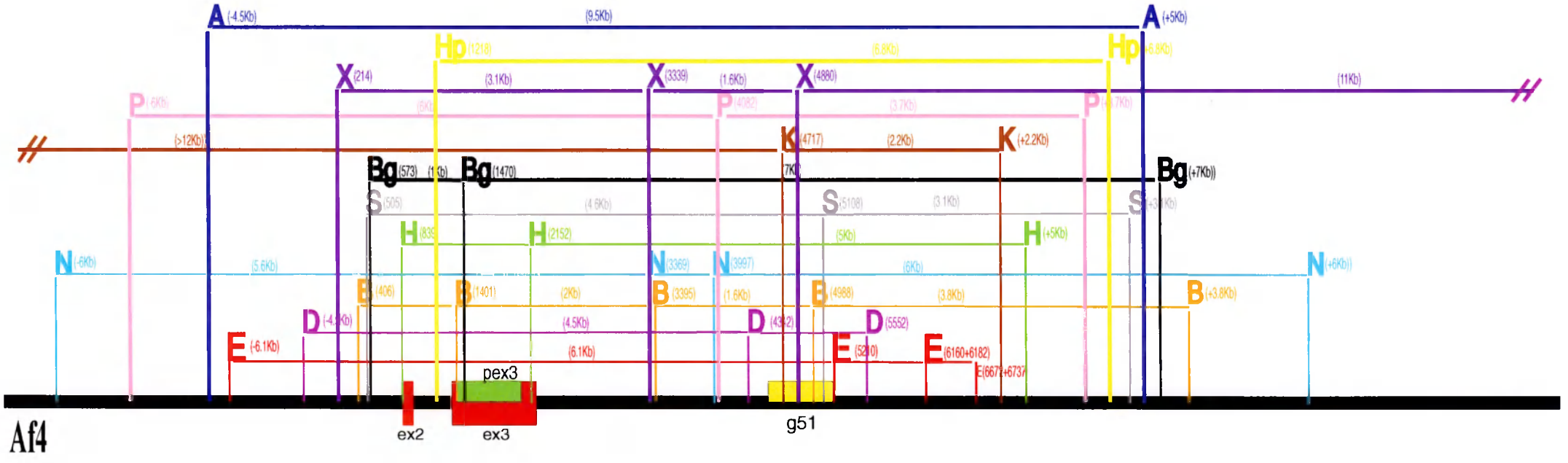


Figure 9 Restriction map of the Af4 BAC

Various sets of single and double restriction digests were used to assemble this map. The two probes used were the exon 3 probe (Pex3) and the probe g51, which hybridises to the third intr
The names of restriction enzymes are abbreviated as follows: AvrII (A), BamHI (B), BglI (Bg), DraI (D), EcoRI (E), HindIII (H), HpaI (Hp), KpnI (K), NheI (N), PvuII (P), SacI (S), XbaI (
Pex3 and g51 are the two probes used for southern hybridisation. See text for further details.

VIII.2.3 Sequencing the region of *Af4* covering the intron 1 through intron 3 interval.

VIII.2.3.1 Direct BAC sequencing

Direct sequencing was performed on the *Af4* BAC around exon 2 and 3 by the EMBL sequencing facility, yielding complete sequence from intron 1 well into intron 3. This information was needed for two reasons: to complement the restriction map required for the design of an adequate Southern strategy to screen ES cells for homologous recombination and to place the loxP flanked selectable marker cassette sA-IRES-βGeok-pA-hygro-SV40 into intron 3.

The following regions were sequenced:

1) 3526 residues upstream of exon 2.

2) The complete sequence of intron 2, encompassing 409 bp.

3) 1145 residues downstream of exon 3. At this site, a region was encountered which proved resistant to any sequencing attempt, because of an extremely high GC content. The only sequencing option would have involved "walking" with sequencing primers from the other side (ie. exon 4), but the inferred length of *AF4* intron 3 in humans (approximately 30 kb) suggested that such an approach would be practically not feasible. This problem highlights two main features of contemporary genomic engineering in higher organisms.

First, despite remarkable progress, the mouse genomic sequence is currently unavailable for easy and direct use in genetic engineering. Furthermore, although DNA sequencing has become a straightforward and high-throughput technique, this example shows that some regions of the genome can be a remarkable hurdle, even with the use of advanced sequencing methods.

Second, current approaches to genome manipulation in higher organisms are progressively moving from traditional cDNA transgenesis techniques (whereby usually simple cDNAs are integrated in the genome randomly), to more precise alterations in the

endogenous locus of a particular gene, which often require working with large inserts of genomic DNA. Hence, large cloning vectors (like BACs) are becoming the tool of choice for higher genomic manipulation. Precise changes in such large genomic fragments require substantial sequence information, and the case of intron 3 of *Af4* was a typical example of a situation in which sequence information was required and not available for a region whose distance from the closest anchor point of known sequence (more than 25 kb to the beginning of exon 4) made any attempt at sequencing by a "primer-walking" method unrealistic.

A variety of approaches can address this kind of problem, including long-range PCR and random subcloning of the BAC with subsequent identification of the positive colonies harbouring the region of interest. Long-range PCR was repeatedly tried in this case without success, indicating that the length and/or the GC content of this locus were beyond the current limits of the technique. I therefore decided to apply a variation of the ET cloning methodology to subclone in one step almost the whole intron 3 from the BAC into a high-copy plasmid . Having this whole region cloned in a conventional plasmid was expected to facilitate direct sequencing and random subcloning of the relevant region. This experiment also established, as a proof of principle, the applicability of ET subcloning to large regions of BACs, which offers obvious advantages in many analogous experimental settings.

VIII.2.3.2 ET recombination can be successfully used to directly subclone large fragments of BACs: application to the third intron of the *Af4* gene.

Oligos *Af4*a and *Af4*b were designed, each containing two regions. The 5' region was identical to the sequence in the BAC immediately flanking the region to be subcloned, and thereby mediated homologous recombination in *E. coli* upon expression of the relevant recombinogenic proteins. The 3' end was a stretch of 40 nucleotides annealing to the plasmid amplified by PCR in whose backbone the target region would be subcloned. EcoRI

and BamHI sites were included in the oligos between the 5' and the 3' ends to facilitate subsequent cloning steps.

Oligo *Af4*b has the following 5'-3' sequence:

TCTCAAAGCTTTGCTGACCATAGTCCTCCAGTGGCAGCTTCAGCTCGGGGG

ACTCACTGGGAACCTGATCCTGACCGTCCATGGGACGGACATATGCCGTGG

ACTTC*GGATCC*TTAGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTA

Residues 1 through 107 (in bold) constitute the arm of homology to the *Af4* BAC (positions 684-790 in *Af4* exon3). A BamHI site (in italics) was inserted at positions 108-113. Residues 114-153 (underlined) constitute the portion of the oligo which anneals to the PCR template pACYC177 at the 3' end of the kanamycin resistance gene.

Oligo *Af4*a has the 5'-3' sequence:

AGCTGAGCCCAGGGGCAAGGCTGCTTTGTACCAGCCTGCTGTCTGCGGGG

GCATCACCTGATGTGGCACTTGCTGTGTGTACAAGGGCTCGCGAGTGACAC

AG*GAATTC*TTAATAAGATGATCTTCTTGAGATCGTTTTGGTCTGCGCG

Residues 1 through 103 (in bold) constitute the arm of homology to the third intron of *Af4* 1105 bp upstream of the start of exon 4. An EcoRI site (in italics) was inserted at positions 111-116. Residues 117-156 (underlined) constitute the portion of the oligo which anneals to the PCR template pACYC177 5' of the replication origin.

As diagramatically shown in figure 10a, the resulting PCR product (*Af4*intronsub) features at its ends the two regions of homology to the *Af4* BAC (donor DNA molecule).

HS996 *E. coli* cells harbouring the *Af4* BAC 189o7 were made competent for heat shock transformation by the Rubidium Chloride method (see Materials and Methods), and were then transformed with the recombinogenic plasmid pBAD-αβγ-Amp, which expresses the recombinogenic proteins Redα and Redβ under the control of an arabinose-inducible promoter. From correct transformants, electrocompetent cells were prepared (see Materials

and Methods) and electroporated with the PCR product *Af4*intronsub, and incubated on kanamycin (25 μg/ml) containing plates. About 160 colonies/plate were observed and 15 of 18 analysed showed an identical EcoRI digestion pattern (figure 10b), consistent with subcloning of a region spanning at least 24 kb (as estimated by electrophoresis).

The correct identity of recombinant clones was confirmed by Southern hybridisation with an oligonucleotide probe spanning 113 bp of sequence located 1144 downstream of the end of exon 3. All candidate colonies displayed the correct hybridisation band, demonstrating that they all contained the desired insert from *Af4* intron 3 (Figure 10b).

5 of the EcoRI restriction fragments were then subcloned in pBluescript vector and completely sequenced. This assembled a sequence region from the third intron spanning 10097 residues downstream of exon 3. Importantly, this approach readily yielded sequence from the region that had proved resistant to all previous sequencing attempts, probably due to the presence of a stretch of 19 Cs.

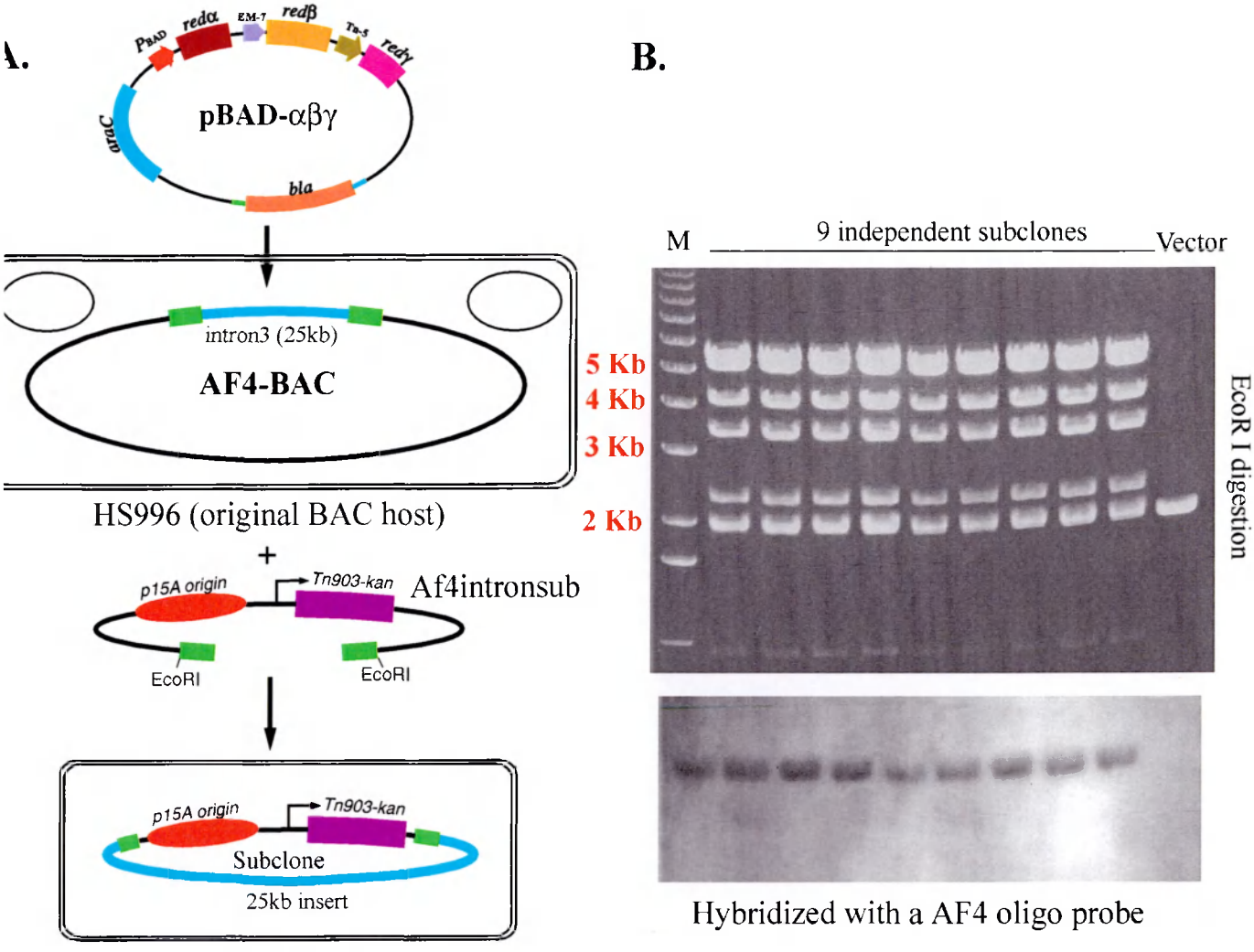


Figure 10 ET mediated subcloning of the third intron of Af4

A. Schematic representation of the ET mediated subcloning of the third intron of Af4. pBADαβγ expresses the recombinogenic proteins Redα Redβ from an Arabinose inducible promoter. The third intron of Af4 was subcloned into a p15A origin plasmid carrying the kanamycin resistance gene. Green blocks represent the arms of homology.

B. 9 independent colonies showed the same restriction pattern, compatible with the length of the subcloned intron as inferred from the human gene. Correct identity of each subclone was confirmed by southern hybridisation with an intron 3 probe.

VIII.2.4 Assembly of the targeting cassette bw-sA-IRES-βGeok-pA-hygro-SV40-aw

VIII.2.4.1 Generation of a universal knock-in targeting vector which is independent of the expression of the target gene in mouse ES cells.

As described above (section VIII.1) I was uncertain about the expression status of *Af4* in mouse ES cells. Hence the βGeok cassette was potentially unsuitable since its ability to deliver G418 resistance relies on placement within an actively transcribed gene. Therefore, I modified this cassette to add the hygromycin phosphotransferase gene, carrying its own promoter (cloned opposite to the direction of transcription of the gene, in order to minimize the risk of possible promoter interference) downstream of the sA-IRES-βGeok-pA component. An outline of the cloning strategy used in the assembly of this cassette is depicted in figures 11 and 12.

To facilitate the introduction of the SV40-hygro cassette into βGeok, I combined an ET cloning based approach with traditional cloning techniques. The SV40-hygro cassette was first subcloned into a recipient plasmid using ET subcloning. Flanking restriction sites, chosen for simplicity of subcloning into βGeok, were included at this step. The SV40-hygro cassette was then subcloned into sA-IRES-βGeok-pA by traditional methodology.

Figure 11 shows the commercially available plasmid pcDNA3.1-hygro, which served as the donor plasmid for the hygromycin phosphotransferase gene. The hygromycin phosphotransferase gene is located under the control of the SV40 promoter. The recipient plasmid was pACYC184, also shown in figure 11, carrying the chloramphenicol resistance gene. To subclone the hygromycin phosphotransferase gene from pcDNA3.1-hygro into pACYC184, a PCR reaction was performed with oligos ShygroF1 and ShygroR, each of which, as in any ET cloning experiment, is composed of two parts, one annealing to the template (in this case pACYC184), the other consisting of around 50 nucleotides of homology to the donor plasmid (in this case pcDNA3.1-hygro) immediately flanking the

target region to be subcloned (in this case the hygromycin phosphotransferase gene). Oligo ShygroF1 has the 5'-3' sequence:

TCGCCGATAGTGGAAACCGACGCCCCAGCACTCGTCCGAGGGCAAAGGAA

TAG*GTTTAAAC*TTAATAAGATGATCTTCTTGAGATCG.

Residues 1 through 53 (in bold) are identical to the region of pcDNA3.1 hygro which covers the 3' end of the hygromycin phosphotransferase gene.

A PmeI site (position 54-61, italics) was inserted to enable the subsequent cloning into PUX4- sA-IRES-βGeok-pA

The 3' end portion (62-87, underlined) anneals in the PCR reaction to pACYC184 template at residues 728-753.

Oligo ShygroR has the 5'-3' sequence:

TCCACACCTGGTTGCTGACTAATTGAGATGCATGCTTTGCATACTTCTGCCT

GCCTG*GGCCGGCCTTCACCGGTGTGCCCGGGCGCGGCCGC*TTACGCCCCGCCCT

GCCACTCATCGC.

Residues 1 through 57 (in bold) are identical to the region of pcDNA3.1-hygro which covers the start of the SV40 promoter.

A polylinker (position 58-92, italics) of unique restriction enzyme sites (FseI, SgrAI, SrfI and NotI) enables both the subsequent cloning into PUX4-sA-IRES-βGeok-pA, and the further insertion of one of the two homology arms for the final ET cloning step into the *Af4* BAC.

The 3' end portion (93-118, underlined) anneals in the PCR reaction to the pACYC184 template.

As depicted in figure 11, the PCR product features at its ends the two regions of homology to the hygromycin phosphotransferase gene. The PCR product was first digested with DpnI, an enzyme which cleaves only when its restriction site (GATC) is methylated, and thereby

serves after the PCR to eliminate the template (in this case pACYC184), which could otherwise contribute a serious source of background in the ET cloning reaction. Following purification of the DpnI digest, the PCR product (300 nanograms) was coelectroporated together with the donor plasmid (pcDNA3.1-hygro) in the E.Coli strain JC8679, which constitutively expresses the recombinogenic proteins RecE and RecT. Cells were then plated on chloramphenicol selection. Eight out of nine colonies displayed the correct restriction pattern consisting of two bands of 2129 and 1133 bp respectively. Subsequently (figure 12), the flanking PmeI and NotI sites were used to subclone into sA-IRES-βGeok-pA, yielding the final plasmid PUX4-sA-IRES-βGeok-pA-hygro-SV40.

Figure 11 ET mediated subcloning of the hygromycin resistance gene
The hygromycin resistance cassette was subcloned from the pcDNA3.1-hygro plasmid into the pACYC184 vector. Convenient restriction sites were incorporated in the ET homology arms to enable subsequent cloning steps.
See text for a full description of the cloning strategy.
Abbreviations: (Cm) chloramphenicol resistance gene; (Amp) βlactamase gene; (ColE1) ColE1 origin of replication; (p15A) p15A origin of replication; (pA) polyadenylation signal; (SV40) simian virus 40.

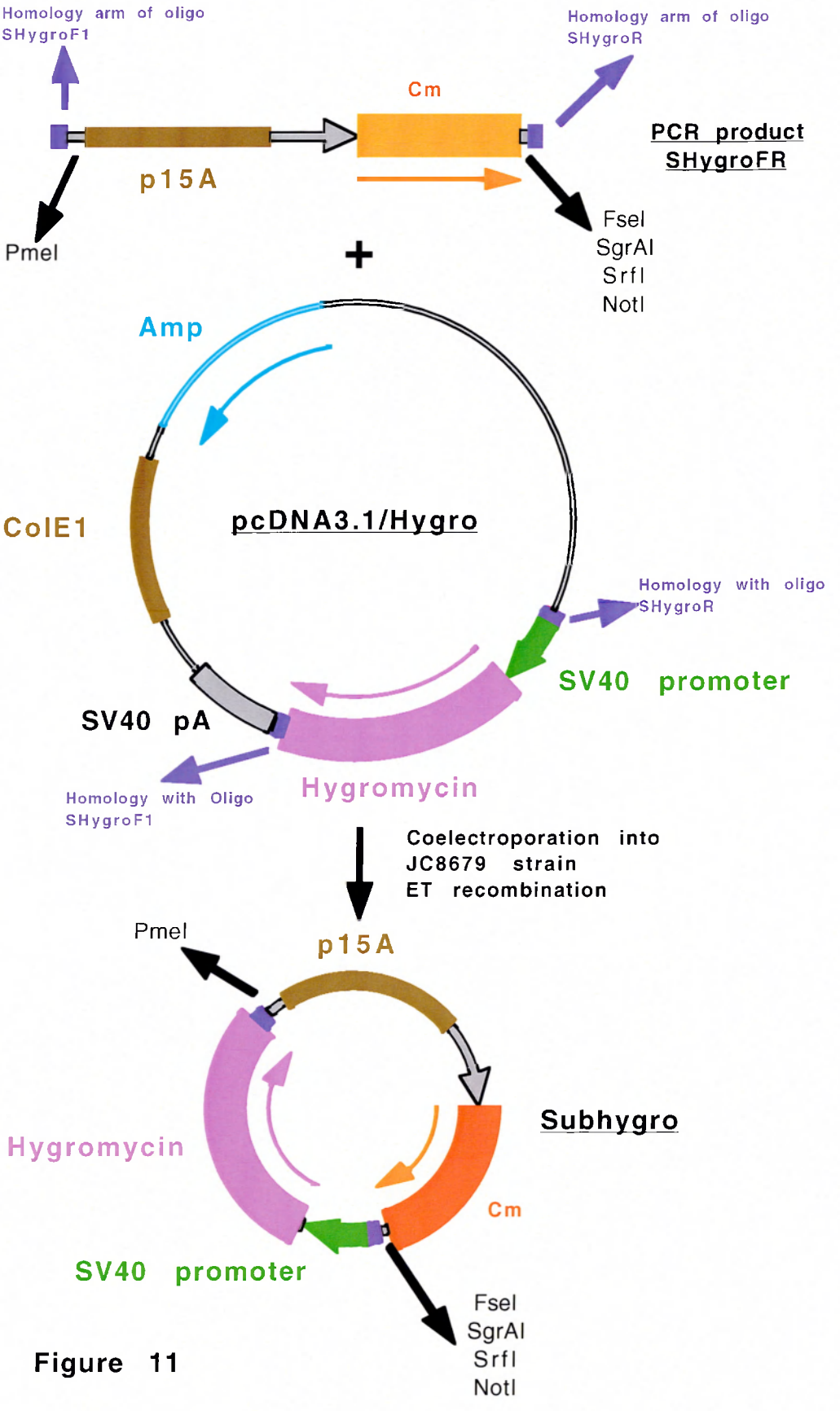

Homology arm of oligo
SHygroF1
Homology arm of oligo
SHygroR
Cm
PCR product
SHygroFR
p15A
PmeI
FseI
SgrAI
SrfI
NotI
Amp
ColE1
pcDNA3.1/Hygro
Homology with oligo
SHygroR
SV40 promoter
SV40 pA
Hygromycin
Homology with Oligo
SHygroF1
Coelectroporation into
JC8679 strain
ET recombination
PmeI
p15A
Subhygro
Hygromycin
Cm
SV40 promoter
FseI
SgrAI
SrfI
NotI

Figure 11

Figure 12 Generation of a universal promoter trap targeting vector.
Schematic illustration of the cloning strategy to generate a universal promoter trap targeting vector based on the βGeok fusion (PUX4-sA-IRES-βGeok-pA-hygro-SV40) and the hygromycin phosphotransferase gene. It can be used to target genes in ES cells, regardless of their expression status, for the generation of mouse lines where the loss of function and the expression of the gene can be studied simultaneously.
The names of the DNA molecules (constructs and PCR products) are in bold and underlined. The scale bar indicates the number of nucleotides/centimeter.
Abbreviations: (Amp) βlactamase gene; (ECMV IRES) encephalomyocarditis virus internal rybosomal entry site; (LacZ-neo) fusion of the β-galactosidase and the neomycin genes (pA) polyadenylation signal; (SV40) simian virus 40.

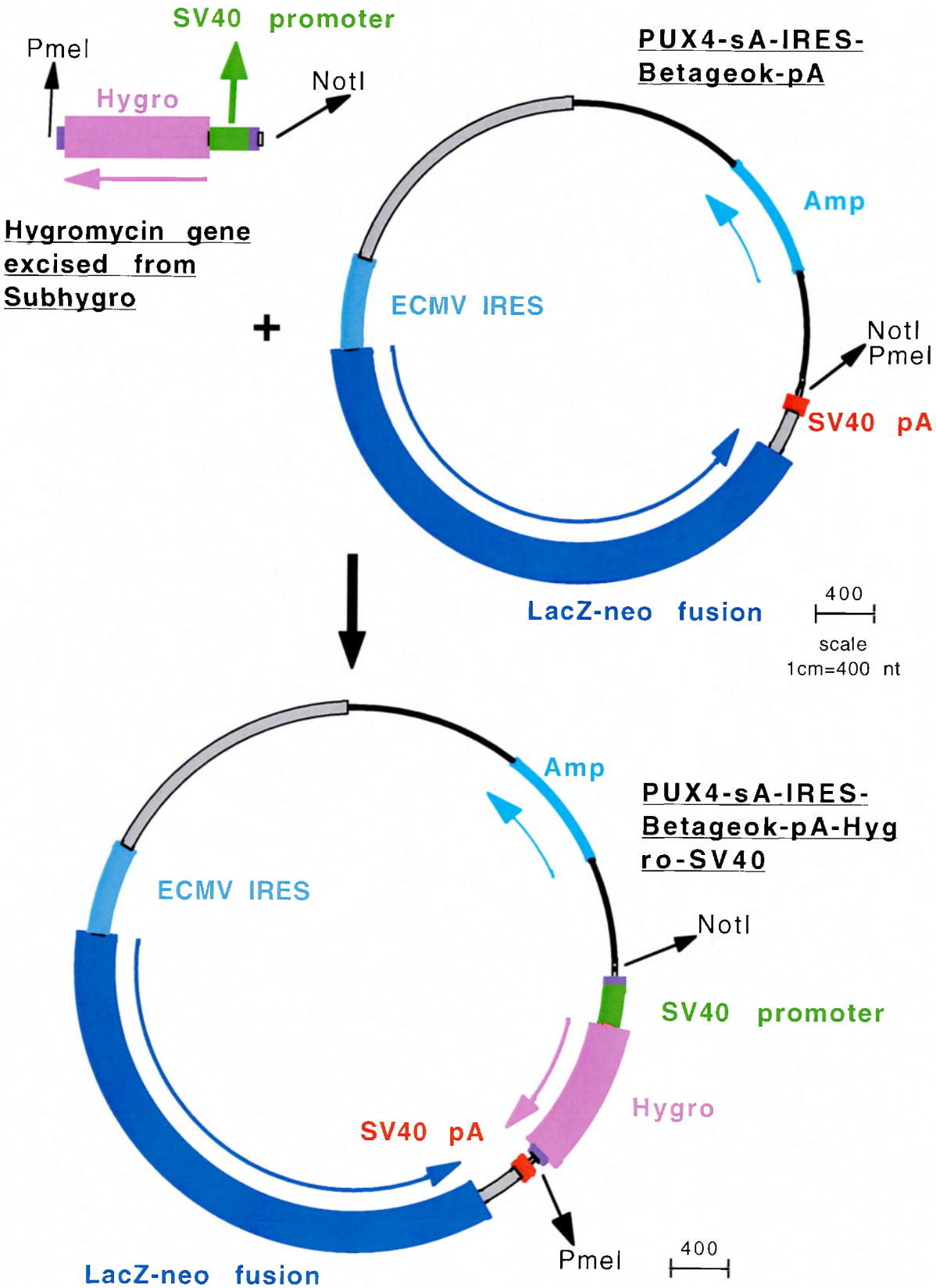

SV40 promoter
PmeI
NotI
Hygro
Hygromycin gene excised from Subhygro
PUX4-sA-IRES-Betageok-pA
Amp
ECMV IRES
NotI
PmeI
SV40 pA
LacZ-neo fusion
400
scale
1cm=400 nt
Amp
PUX4-sA-IRES-Betageok-pA-Hygro-SV40
ECMV IRES
NotI
SV40 promoter
Hygro
SV40 pA
PmeI
LacZ-neo fusion
400
scale
1cm=400 nt

Figure 12

VIII.2.4.2 Assembly of the final ET targeting cassette bw-loxP-sA-IRES-βGeok-pA-hygro-SV40-loxP-aw

As already described, ET cloning normally utilises a PCR reaction to incorporate at the two ends of a given donor DNA molecule the two regions of homology which will direct its incorporation at the desired site into the acceptor DNA molecule. Given the possibilities offered by current long-range PCR protocols, this approach could also be considered in this case, amplifying the newly generated sA-IRES-βGeok-pA-hygro-SV40 with oligonucleotides which add the homology arms for ET recombination. However, the inherent risk of mutations generated during the PCR reaction over long distances discourages this approach. Hence, a combination of ET cloning and conventional cloning techniques represented the best solution to this problem. In the first step, short homology arms were made by PCR and placed by conventional cloning at the two ends of the targeting vector, using unique restriction sites incorporated in the early steps of the cloning plan. This assured that the core of the targeting construct was never subjected to PCR amplification. The homology arms could have been directly synthethized as oligonucleotides and added by ligation, however the cheaper alternative of PCR generation was preferred.

The *Af4* BAC 189o7 was used as a template to generate the 5' and 3' homology arms, called bw and aw respectively (diagram in Figure 13). The region to be targeted in intron 3 was chosen after an analysis of the sequence available. First, current literature and the Genebank database were searched for the *AF4* genomic breakpoint sequences in human patients. Unfortunately, little sequence information is available. It would have been interesting to compare such sequences with the mouse intron 3 sequences. Even in the absence of significant homology (a rather likely outcome for intronic sequences), it would have been valuable to place the loxP site in the mouse intron at roughly the same distance from exon 3 or 4 as in the human cases. Unfortunately, even for *MLL-AF4* leukemia cases analysed by

Figure 13 Generation of the ET targeting cassette for the Af4 BAC

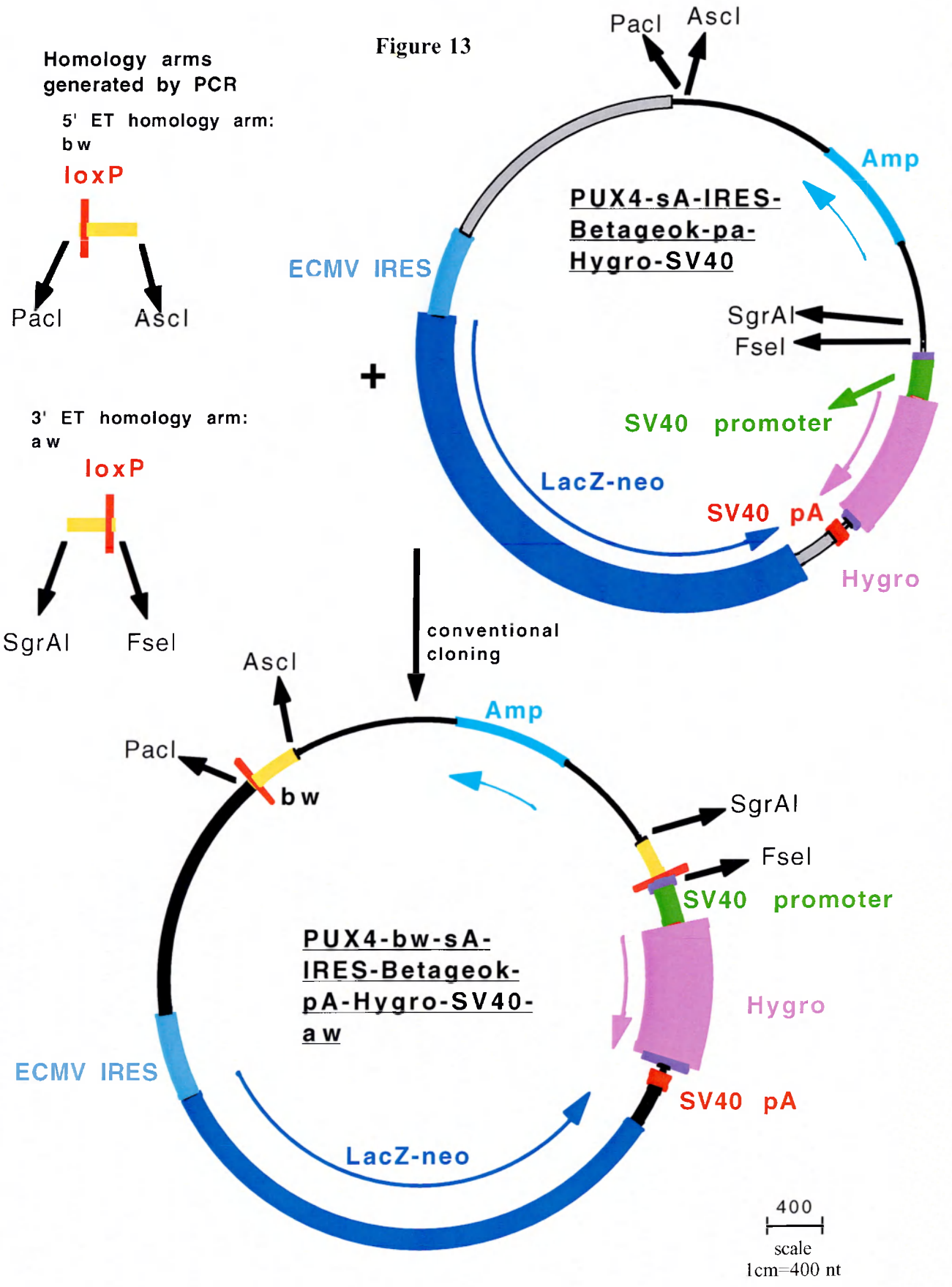
The figure shows the cloning steps involved in the generation of the ET targeting construct PUX4-bw-sA-IRES-βGeok-pA-hygro-SV40-aw.

See text for a full description of the cloning strategy.

The names of the DNA molecules (constructs and PCR products) are in bold and underlined. The scale bar indicates the number of nucleotides/centimeter.

Abbreviations: (Amp) βlactamase gene; (ECMV IRES) encephalomyocarditis virus internal rybosomal entry site; (LacZ-neo) fusion of the β-galactosidase and the neomycin genes (pA) polyadenylation signal; (SV40) simian virus 40; (bw) 5' ET homology arm; (aw) 3' ET homology arm.

Figure 13


Homology arms generated by PCR
5' ET homology arm: b w
loxP
PacI
AscI
3' ET homology arm: a w
loxP
SgrAI
FseI
PacI
AscI
Amp
PUX4-sA-IRES-Betageok-pa-Hygro-SV40
ECMV IRES
SgrAI
FseI
SV40 promoter
LacZ-neo
SV40 pA
Hygro
conventional cloning
AscI
PacI
b w
Amp
SgrAI
FseI
SV40 promoter
PUX4-bw-sA-IRES-Betageok-pA-Hygro-SV40-aw
Hygro
ECMV IRES
SV40 pA
LacZ-neo
400
scale
1cm=400 nt

PCR or Southern blotting, the actual intronic breakpoints had not been sequenced at the time of these experiments.

Second, on the basis of the sequence available and the mapping information, the disposition of NheI sites presented a good Southern strategy to screen the ES cell clones for homologous recombination. Two NheI sites in the 5' end of intron 3 were identified, whose replacement by the targeting construct generates bands of different sizes in the wild type and homologously recombined alleles. The two NheI sites flank the stretch of 19 Cs, which confounded the sequence characterisation of the BAC, and a B1 repeat. B1 repeats are the most abundant type of repeats in the mouse genome, and they have so far not been associated with any function. Hence, it was decided to design the ET cloning homology arms in such a way that the loxP-sA-IRES-βGeok-pA-hygro-SV40-loxP targeting vector would, upon homologous recombination in the *Af4* BAC, replace the 700 bp region between the two NheI sites, including the 19 Cs and the B1 repeat. The 5' homology arm was amplified by PCR from BAC DNA using the oligos N*af4*bwF and N*af4*bwR. Forward oligo N*af4*bwF has the 5'-3' sequence:

TTACATT*GGCGCGCC*TAATTATCCTGAACCCTCCCACCTC.

At position 8-15, an AscI (italics) site was incorporated for further cloning, while the annealing portion includes residues 16 to 40. Reverse oligo N*af4*bwR has the 5'-3' sequence:

CACGGTGCC*TTAATTAA***ATAACTTCGTATAGCATACATTATACGAAGTTATAG**G CTTTCTCACACGCAATACAGGT.

At position 10-17, a PacI site (italics) was incorporated for futher cloning; the loxP site (in bold) is located at position 18-51, while the annealing portion (underlined) includes residues 52-76. The resulting PCR product (bw) is 403 nucleotides long and contains 336 nucleotides of sequence identity with the region of the BAC to be targeted.

The 3' homology arm was amplified from *Af4* BAC DNA using the oligos N*af4*awF2 and N*af4*awR. Forward oligo N*af4*awF2 has the 5'-3' sequence:

TTACCATTCA*GGCCGGCC***ATAACTTCGTATAATGTATGCTATACGAAGTTAT**G**

AGGGAGAGCCAATCAGAAGA.

An FseI site (position 11-18, italic) was incorporated for further cloning. The loxP site (in bold) is located at position 19-52, while the annealing portion includes residues 53-73 (underlined). Reverse oligo N*af4*awR has the 5'-3' sequence:

CTAATGGTAA*CACCGGTG*GCTACCAAGACTGACAAACCAAG.

An SgrAI site (position 11-18, italic) was inserted for further cloning. The annealing portion (underlined) includes residues 19-41. The resulting PCR product (aw) is 336 base pairs long and contains 265 residues of sequence identity with the region of the BAC to be targeted.

The final product was the bw-loxP-sA-IRES-βGeok-pA-hygro-SV40-loxP-aw cassette, ready to be placed by ET recombination into the backbone for the targeting construct subcloned from the BAC.

VIII.2.5 ET-mediated subcloning of the *Af4* target region from the *Af4* BAC into the vector pACYC-177

To assemble the final *Af4* targeting construct for homologous recombination in mouse ES cells, two sequential steps of ET recombination were applied. The first step is described here, and the second in section VIII.2.6.

A subclone from the BAC was made, by ET recombination, to establish the backbone of the targeting vector including the 5' and 3' homology arms for use in ES cells. These homology arms were chosen by design so that the boundaries of the final ES targeting vector suited a convenient Southern strategy. The *Af4* targeting construct described below exemplifies the merits of this strategy based on NheI, since it could be used to screen for

homologous integration at both the 5' and the 3' ends. Furthermore, the finite amount of DNA available for each ES clone is an important consideration for any Southern strategy, limiting the number of digests which can be performed. Use of a single digest, and hence a single blot, renders the screening strategy easier, faster, and less prone to errors in ordering samples and interpreting the results.

Another advantage of using ET recombination to establish the ES cell targeting vector backbone exploited in this *Af4* example, involves the easy placement of a chosen, unique restriction site flanking the subclone. Subsequent cleavage of these flanking sites yields the linear fragment for targeting in ES cells.

To subclone the *Af4* targeting backbone from the *Af4* BAC into pACYC177 (Figure 14), a PCR reaction was performed with oligos *Af4*5End and N*af4*3End, each of which, as in any ET cloning experiment, is composed of two parts: the PCR primer annealing to the template (in this case pACYC177), and the homology arm to the donor plasmid (in this case the *Af4* BAC) that defines the target region to be subcloned. Forward oligo *Af4*5End has the following 5'-3' sequence:

ACAGGCATTTGTCAACTACACCCACCGGGCAGGCTCCAAGTTCACTCACTA CAAATCTGTAGTAGCTGAAGTGTTCTTG*GCGGCCGC*TCACGAGGCAGACCTC AGCGCTAGCGG.

Residues 1-79 constitute the 5' ET homology arm. A NotI site was incorporated (position 80-87) to release the insert from the targeting vector prior to ES cell electroporation. The annealing part to the pACYC177 template includes residues 88 to 114. Reverse oligo N*af4*3End has the following 5'-3' sequence:

GCCCTAGGAGCTAGTTCTCAGTGAATGTATGTTACACAAGTATAGTTCCTGT AACCCCTGCATAGGGCATGGGGAT*GCGGCCGC*TGAAGACGAAAGGGCCTCGT GATACGCC.

Residues 1-76 constitue the 3' ET homology arm. Also in this oligo, a second NotI site was incorporated (position77-84) to release the insert from the targeting vector prior to ES cell electroporation. The annealing part to the pACYC177 template includes residues 85 to 112. As shown in figure 14 the PCR product (pACYC177*Af4*) features at its ends the two regions of homology to the *Af4* BAC.

E. coli cells harbouring the *Af4* BAC 189o7 were transformed with the recombinogenic plasmid R6K-αβγ-Tet, which expresses the recombinogenic proteins Redα and Redβ under the control of an arabinose-inducible promoter. The R6K origin of replication allows this plasmid to coexist in the same cell with ColE1 origin plasmids, like pACYC177, a prerequisite for this kind of experiment where the presence of two plasmids is necessary. Electrocompetent cells were electroporated with the PCR product pACYC177*Af4*., and spread on ampicillin (100 μg/ml) plates. On each plate, approximately 250 colonies arose and twelve colonies were picked from different plates. In 4/12 colonies the expected pattern was detected (3 bands of respectively 6815, 4544 and 623 base pairs) (Figure 15). Since this result was less efficient than the previous subcloning of a larger section of the same region from the same BAC, positive candidates were then digested with the following battery of restriction enzymes (ScaI, NotI, KpnI, ApaI, NheI, and SacI) to further confirm that they contained the correct fragment (figure 16). The expected patterns were observed (for ScaI Three bands of 5478, 3584 and 2929 base pairs; for NotI, two bands of 9924 and 2067; for KpnI, two bands of 9729 and 2262 base pairs; for ApaI, a single band of 11991 base pairs, corresponding to the linearised plasmid; for NheI three bands of 6815, 4544 and 623 base pairs; and for SacI, four bands of 4608, 3287, 3224 and 872 base pairs). Colony n.2 (*Af4*sub2) was then linearised with ApaI, since it was shown that the ET cloning reaction is more efficient with a linearised than a supercoiled DNA target. In this case, the ApaI site (position 6927-6932 of the *Af4*Sub construct) was chosen since it is located within

the region of *Af4* to be replaced by the bw-loxP-sA-IRES-βGeok-pA-hygro-SV40-loxP-aw cassette.

Figure 14 ET mediated subcloning of the *Af4* target region from the Af4 BAC into the vector pACYC177

The relevant region for the ES targeting construct was subcloned via ET recombination from the *Af4* BAC into the middle-copy vector pACYC177.

See text for a full description of the cloning strategy.

The names of the DNA molecules (constructs and PCR products) are in bold and underlined. The scale bar indicates the number of nucleotides/centimeter. Due to the size of the *Af4* BAC molecule, the dashed line region of the Af4 BAC construct represents a region greater than 100 Kb.

Abbreviations: (Cm) chloramphenicol resistance gene; (p15A) p15A origin of replication; (pA) polyadenylation signal; (SV40) simian virus 40.

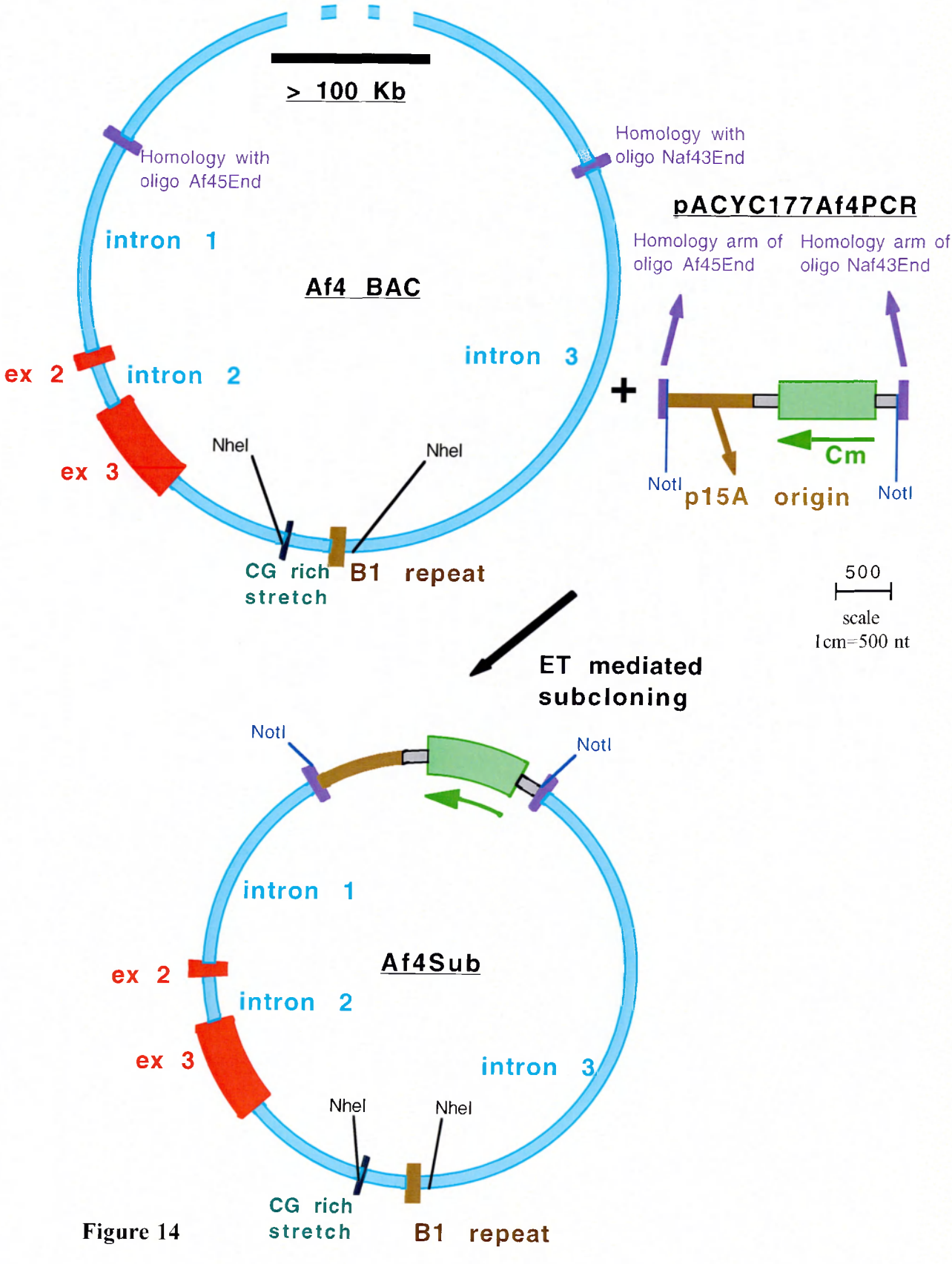

> 100 Kb
Homology with oligo Af45End
Homology with oligo Naf43End
intron 1
Af4 BAC
intron 3
ex 2
intron 2
ex 3
NheI
NheI
CG rich stretch
B1 repeat
pACYC177Af4PCR
Homology arm of oligo Af45End
Homology arm of oligo Naf43End
NotI
p15A origin
Cm
NotI
500
scale
1cm=500 nt
ET mediated subcloning
NotI
NotI
intron 1
Af4Sub
ex 2
intron 2
ex 3
intron 3
NheI
NheI
CG rich stretch
B1 repeat

Figure 14

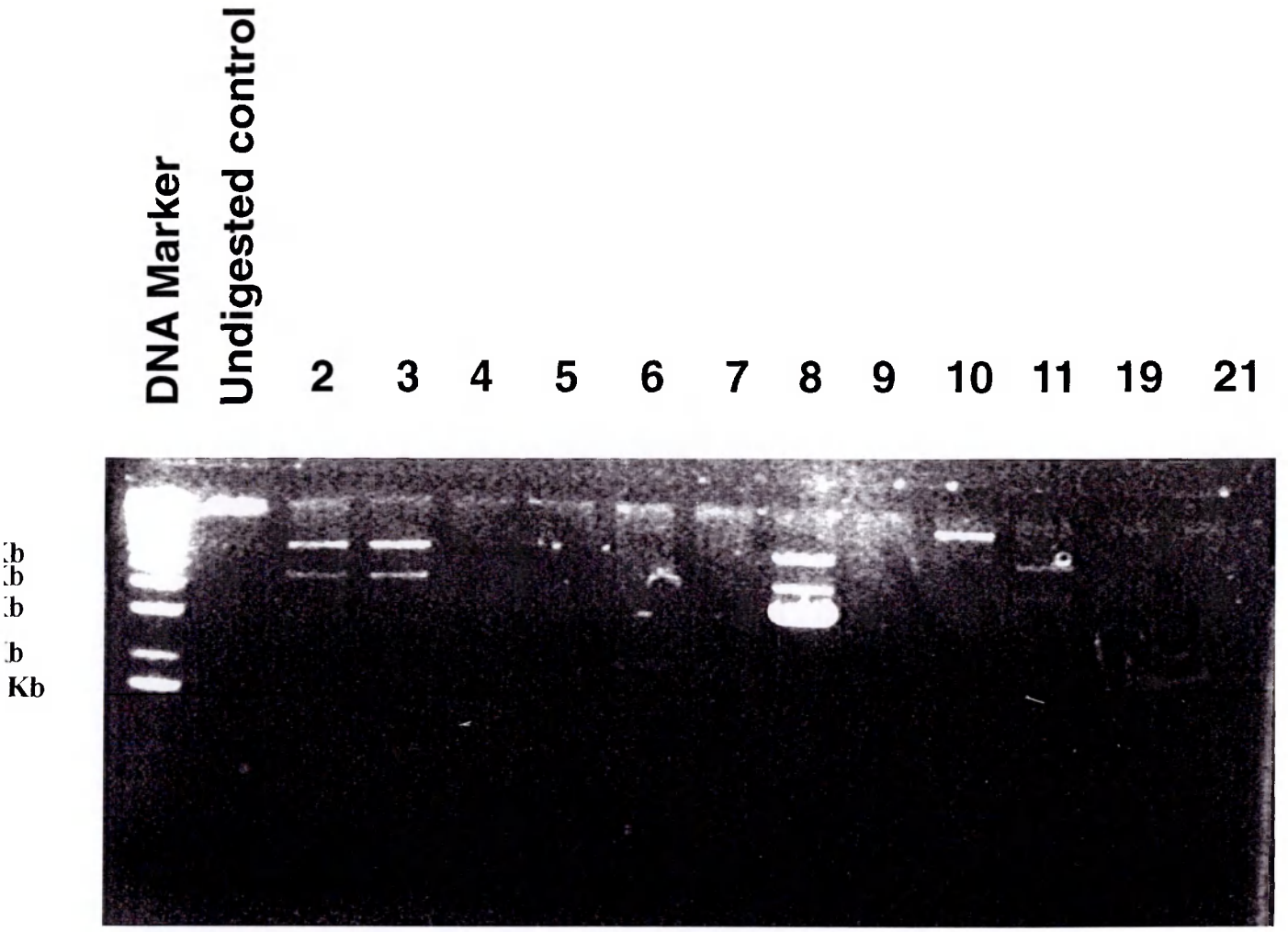


Figure 15 ET mediated subcloning of the Af4 targeting backbone
NheI digests of 12 colonies resulting from the subcloning of a fragment of the Af4 BAC into the vector pACYC177 to yield the backbone for the assembly of the knock-in targeting construct. Colonies 2, 3, 8 and 11 show the correct restriction pattern.

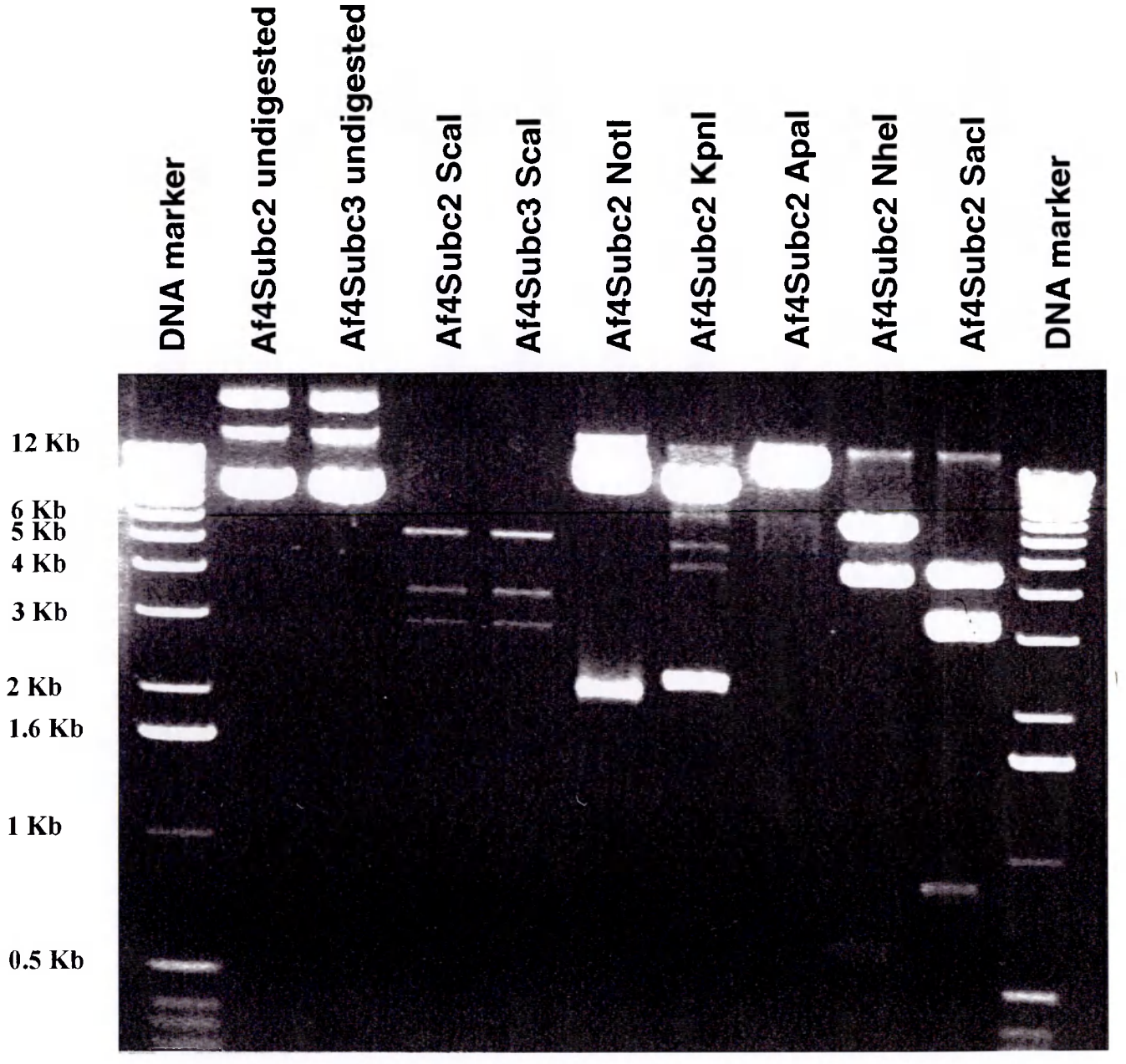


Figure 16 Restriction analysis of the Af4 subclone
Restriction digest of the fragment of the Af4 BAC subcloned into the pACYC177 vector (Af4Sub) as a starting backbone for the assembly of the knock-in targeting construct. Af4Sub colonies n.2 and n.3 were digested with ScaI, NotI, KpnI, ApaI, NheI and SacI to check for the correct restriction pattern.

VIII.2.6 Insertion of the bw-loxP-sA-IRES-βGeok-pA-hygro-SV40-loxP-aw cassette into the *Af4* subclone

In the final step to assemble the *Af4* ES cell targeting construct (*Af4*-loxP-βGeok), the bw-loxP-sA-IRES-βGeo-pA-hygro-SV40-loxP-aw cassette was cloned into the *Af4* target backbone, *Af4*Sub, by ET recombination (Figure 17).

PUX4-bw-loxP-sA-IRES-βGeo-pA-hygro-SV40-loxP-aw was digested with NotI and AscI, to release the recombinogenic bw-loxP-sA-IRES-βGeo-pA-hygro-SV40-loxP-aw insert. The released insert was gel purified and cotransformed with the ApaI linearised *Af4*Sub into two different *E. coli* strains: DH10β which had been transformed with the recombinogenic plasmid R6Kαβγ; and JC8679, which constitutively express the recombinogenic proteins RecE and RecT. Successful ET recombinants were identified by growth on double selection plates (kanamycin 50 μg/ml and ampicillin 100 μg/ml). A total of 20 colonies, 16 from the DH10β strain, and 4 from the JC8679, were analysed by ScaI digestion (figure 18a). All 20 colonies showed the correct pattern (4 bands of 7594, 5478, 3584 and 3042 base pairs.) Colony 2 was chosen for a more detailed restriction analysis in order to confirm that the ET cloning reaction had indeed generated the correct plasmid (figure 18b).

Figure 17 ET mediated targeting of the loxP-βGeok-hygro cassette to the Af4 subclone

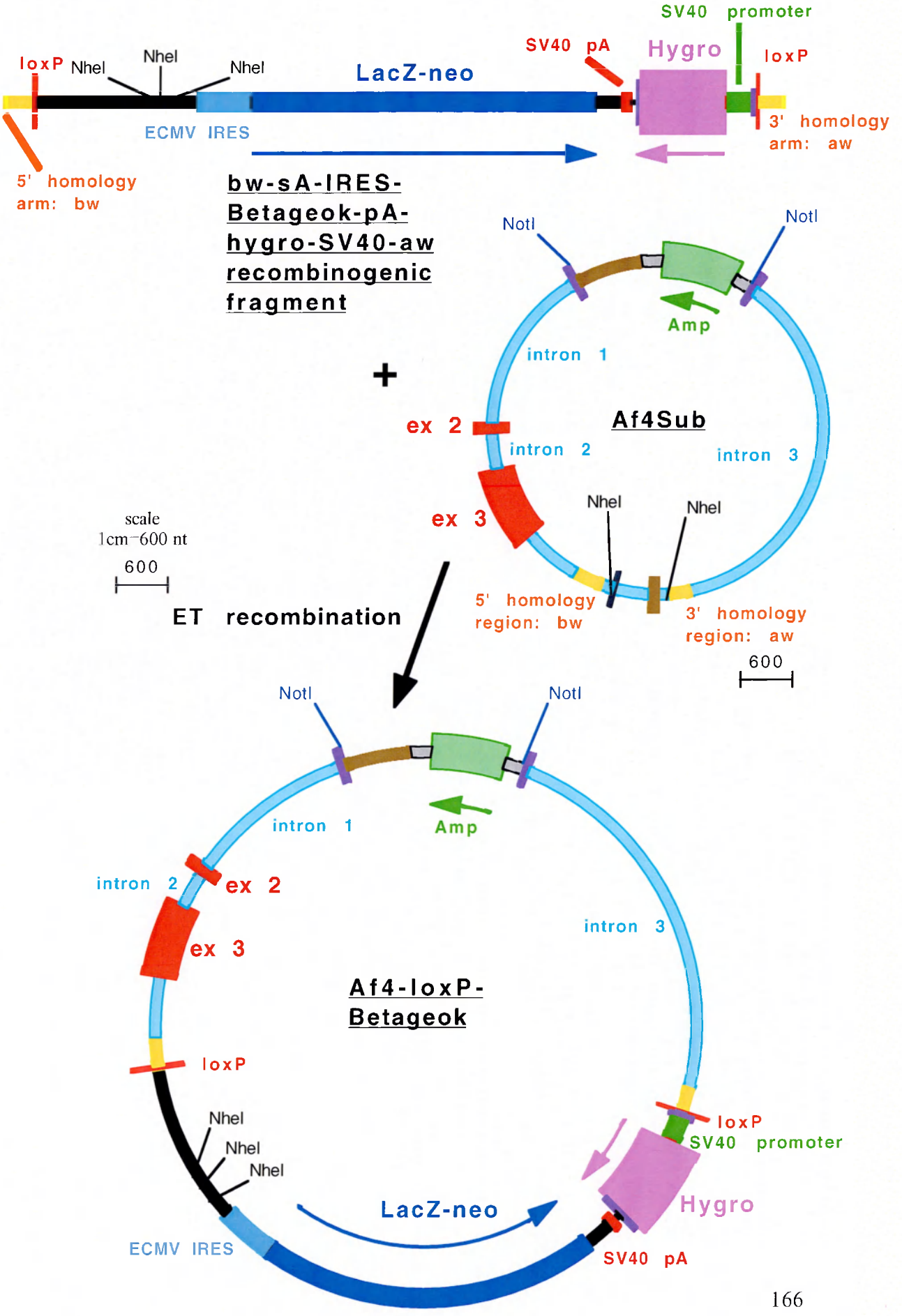
The bw-loxP-sA-IRES-βGeok-pA-hygro-SV40-loxP-aw cassette was targeted by ET mediated homologous recombination to the Af4 targeting backbone, which had been previously subcloned into the pACYC177 vector (Af4Sub).

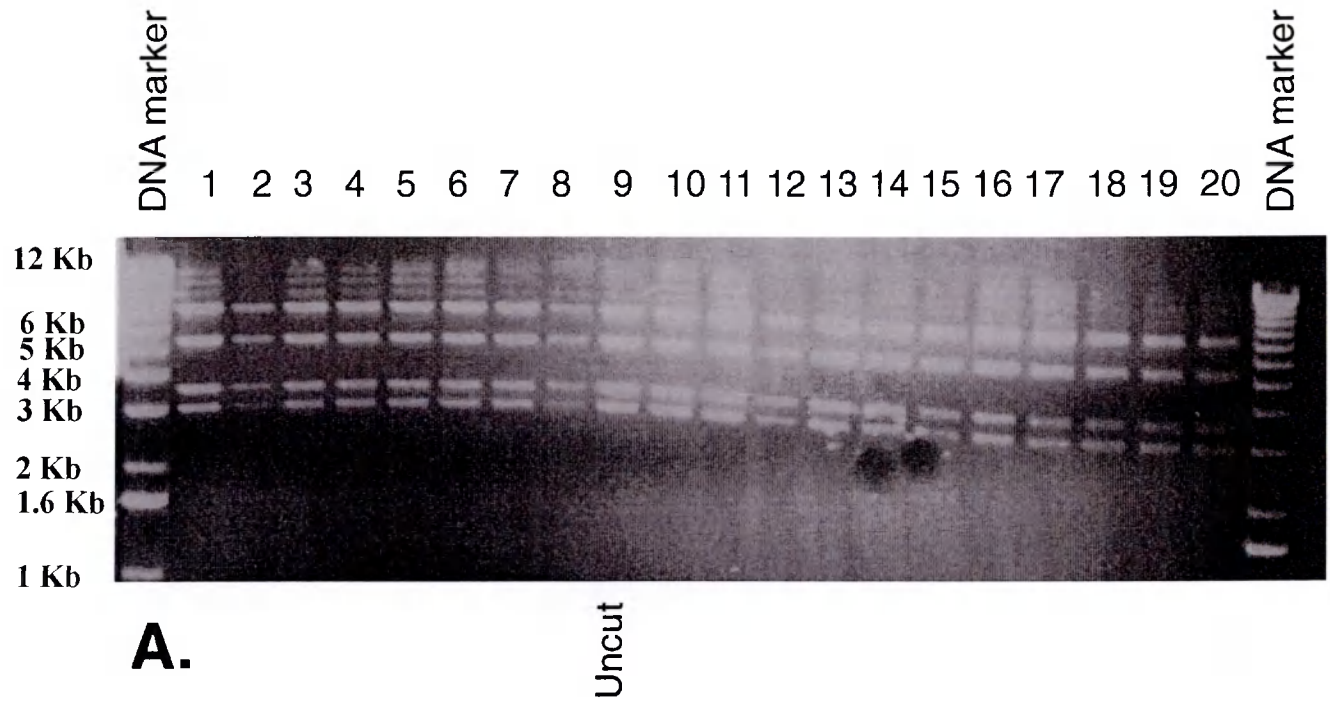
The names of the DNA molecules (constructs and PCR products) are in bold and underlined. The scale bar indicates the number of nucleotides/centimeter.

Abbreviations: (Cm) chloramphenicol resistance gene; (ECMV IRES) encephalomyocarditis virus internal rybosomal entry site; (LacZ-neo) fusion of the β-galactosidase and the neomycin genes (pA) polyadenylation signal; (sA) splice acceptor; (SV40) simian virus 40; (bw) 5' ET homology arm; (aw) 3' ET homology arm.

See text for a full description of the cloning strategy.

Figure 17


SV40 promoter
SV40 pA
Hygro
loxP
loxP
NheI
NheI
NheI
LacZ-neo
ECMV IRES
3' homology arm: aw
5' homology arm: bw
bw-sA-IRES-Betageok-pA-hygro-SV40-aw recombinogenic fragment
NotI
NotI
Amp
intron 1
Af4Sub
ex 2
intron 2
intron 3
ex 3
NheI
NheI
scale
1cm=600 nt
600
ET recombination
5' homology region: bw
3' homology region: aw
600
NotI
NotI
intron 1
Amp
intron 2
ex 2
ex 3
intron 3
Af4-loxP-Betageok
loxP
NheI
NheI
NheI
loxP
SV40 promoter
LacZ-neo
Hygro
ECMV IRES
SV40 pA



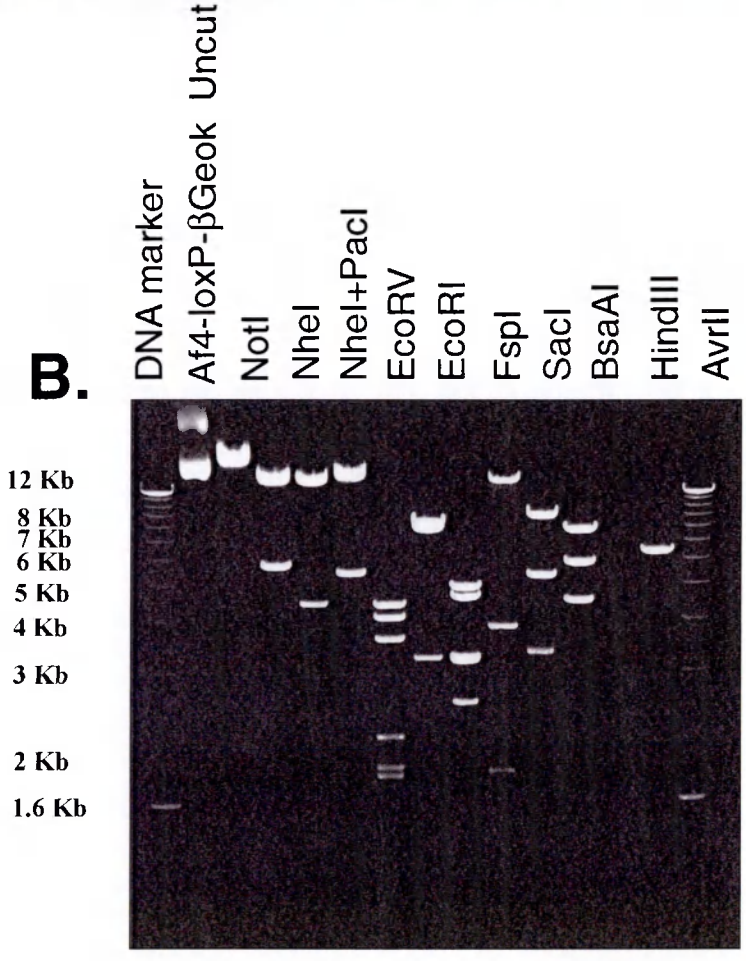


Figure 18 ET mediated targeting of the loxP-βGeok-hygro cassette to the Af4 subclone (restriction analysis)

The loxP-flanked βGeok-hygro cassette was targeted to the third intron of the Af4 gene previously subcloned in pACYC177.

A. All colonies show the correct pattern upon ScaI digestion

B. Colony n. 2 (Af4-koxP-βGeok) is checked with a battery of restriction enzymes, which yield the expected digestion patterns

VIII.3 ES cell targeting with the *Af4* construct

Mouse E14 ES cells (from the line originated in K. Rajewski's lab) were used for the targeting experiment. They were grown on confluent layers of inactivated mouse embryonic fibroblasts (MEFs). The construct *Af4*-loxP-βGeok colony n.2 (40 μg) was digested with NotI to release the targeting construct from the pACYC177 vector. The digested DNA was precipitated and resuspended in 50 μl PBS. On the day of the electroporation, 10 X 10^7 ES cells were trypsinized, resuspended in 700 μl PBS and electroporated with the 40 μg of construct at 240 Volts. After electroporation, cells were plated at low-middle density onto 8 10 cm dishes.

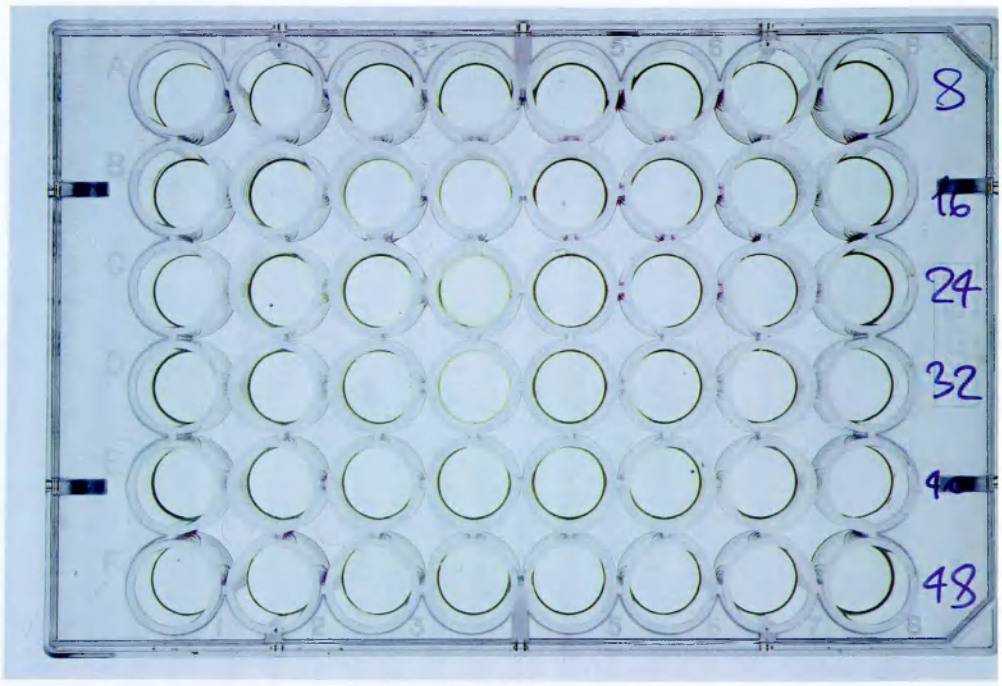
As discussed above, the expression status of the *Af4* gene in mouse ES cells was unclear. Therefore, it was decided to divide the 8 dishes into two sets, one which would be selected with G418, and the other with hygromycin. G418 resistance is conferred by the neo part of the βGeok fusion protein, and, in the configuration of the of *Af4*-loxP-βGeok construct, it relies upon the endogenous expression of the targeted gene. In contrast, hygromycin resistance is autonomous since it is expressed from the SV40 promoter. Since expression of G418 resistance relies on a promoter trap, it seemed likely that selection with G418 would result in much less colonies than with hygromycin. Moreover, if *Af4* was expressed in ES cells, one would predict a much higher frequency of homologously targeted colonies among the G418 than the hygromycin selected ones, since only in the former selection scheme would the promoter trap nature of the construct have been advantageously exploited.

Selection was started 1 day, in the case of G418, and 2 days, in the case of hygromycin, after electroporation, with the following drug concentrations in the cell culture medium: G418 (250μg/ml) and hygromycin (200 μg/ml). Selective medium was then

changed daily. G418 resistant and hygromycin resistant MEFs were used during selection. Nine days after the start of selection, massive cell death had occurred in all plates treated, and surviving ES cells had formed fully grown colonies, which were then picked and transferred into drug resistant MEF coated 48-well plates. Unexpectedly there were comparable number of colonies in either selection scheme. In total, 96 hygromycin resistant and 48 G418 resistant colonies were picked. After four days, each 48-well plate was trypsinised and replica plated into two 24-well plates and two additional 48-well plates. Colonies in the 24-well plates were grown to confluence for DNA extraction. Of the two additional 48-well plates, one was frozen for future reference, while the other one was grown to confluence and stained with X-Gal to assess the expression of βgalactosidase. All plates were scored to assess the percentage of colonies effectively surviving after picking, as well as the morphology of the colonies as a partial indicator of their undifferentiated state.

Figure 19 shows the X-Gal stained replica plates of 48 hygromycin resistant and 48 G418 resistant colonies. There were many more blue colonies among the G418 resistant than among the hygromycin resistant ones, hardly a surprise since G418 resistance and β-galactosidase activity are linked in the same fusion protein βgeo. However, it is known that the sensitivity of detection of both activities (G418 resistance and X-gal staining) can vary, and that cells which are clearly resistant to G418 may not necessarily show detectable X-Gal staining. From the staining of figure 19, this appears to be clearly the case for the following colonies: 10, 18, 21, 41 and 44. On the contrary, lack of staining in the wells 2, 11, 13, 36, 37, 40 and 48 reflects the absence of colonies or an exceedingly low number of cells in these wells. It has been estimated that promoter-trap vectors provide a 100-fold enrichment for homologously targeted versus randomly integrated clones. Moreover, though ideal for genes expressed at high levels in ES cells, promoter trap selection schemes have also been successfully applied to poorly expressed genes.

Hygromycin Selection 200 microgr/ml



G418 Selection 250 microgr/ml

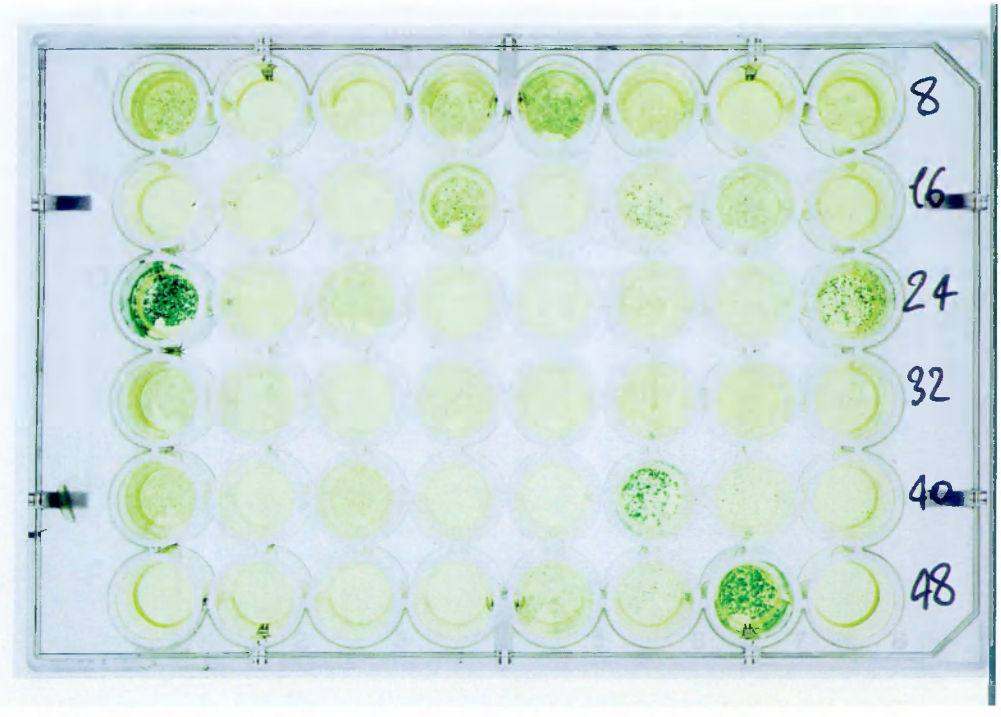


Figure 19 LacZ staining of ES cells targeted with the Af4 construct
The majority of clones selected with G418 versus only few of those selected with hygromycin are positive for LacZ staining, reflecting the architecture of the targeting construct in which the hygromycin resistance gene is expressed from its own promoter, while the βGeok fusion is a promoter trap.
See text for details.

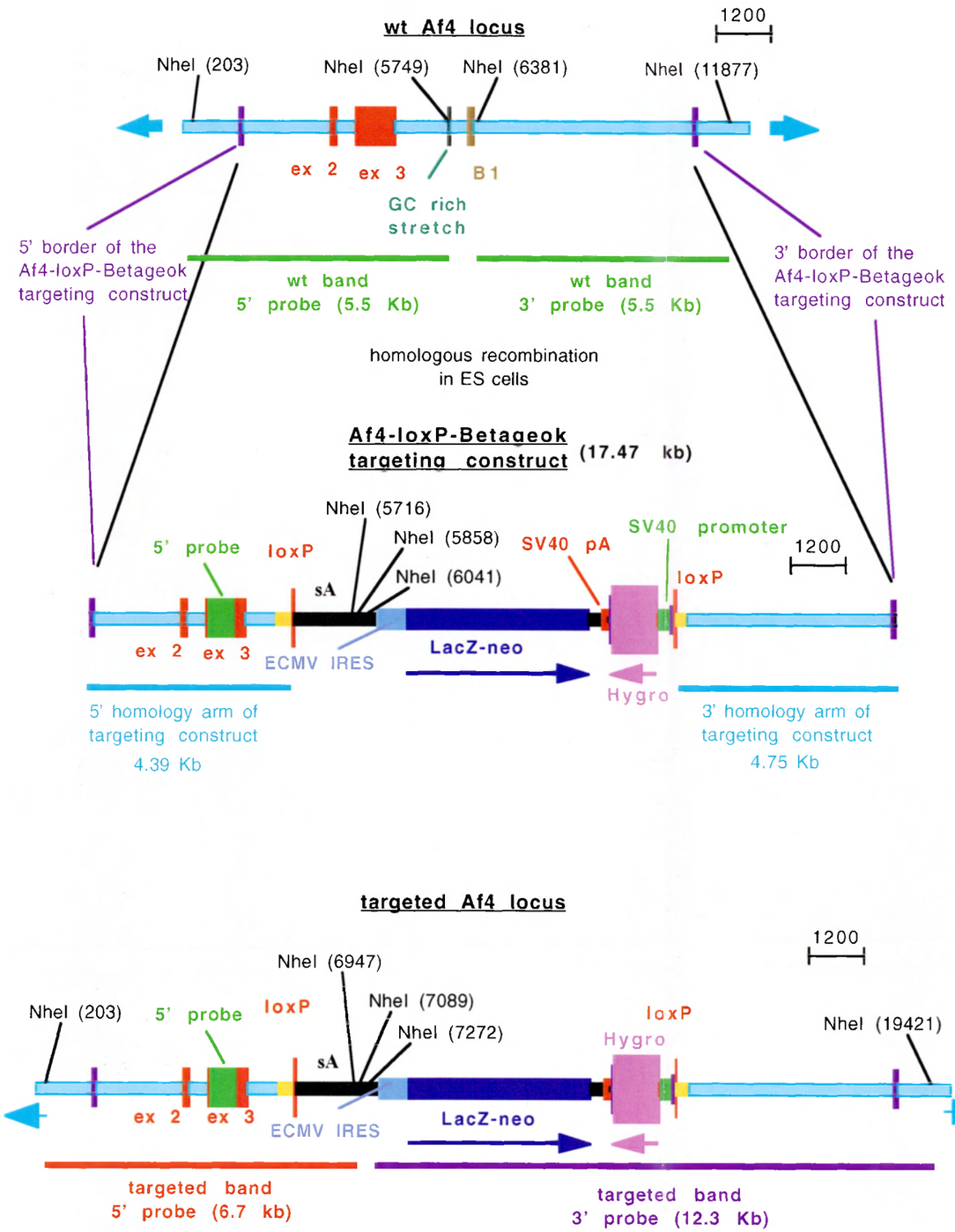
VIII.3.1 Southern blot analysis of G418 resistant colonies

Genomic DNA was extracted from the 48 G418 and 96 hygromycin resistant ES cell colonies and digested with NheI for analysis by Southern blotting with the probes indicated in figure 20.

To screen for homologously recombined colonies on the 5' side of the *Af4* locus, a probe was used (Probe-ex3) which spans 617 nucleotides of *Af4* exon 3. On a genomic NheI digest blot, it hybridises to a wild type band of 5546 bp, resulting from NheI cleavage at positions 508 and 6054 of the wild type *Af4* allele (*Af4*BACwt). In the case of correct integration of the the *Af4*-loxP-βGeok construct in the endogenous *Af4* locus, the NheI site at position 6054 would be lost, since it had already been deleted during the ET cloning of the selectable marker cassette bw-loxP-sA-IRES-βGeok-pA-hygro-SV40-loxP-aw into the *Af4* fragment subcloned into pACYC177 (*Af4*Sub). However, the cassette bw-loxP-sA-IRES-βGeok-pA-hygro-SV40-loxP-aw contains three NheI sites , and therefore, upon NheI digest, correctly targeted colonies would show a new band of 6744 bp, resulting from NheI cleavage at positions 508 and 7252 of the correctly targeted *Af4* allele.

To screen for homologous recombined colonies on the 3' side of the *Af4* locus, a probe was used (neo), which is an EcoRI fragment of the PUX4-βGeok plasmid. In the case of correct integration of the *Af4*-loxP-βGeok construct in the endogenous *Af4* locus, upon NheI digestion, the "neo" probe hybridises to a band of 12312 bp, resulting from NheI cleavage at positions 7577 and 19889 of the correctly targeted *Af4* allele. An advantage of using at least one probe internal to the targeting construct is the ability to distinguish single-copy from multiple-copy ES cell clones, since multiple integrant ES clones cannot be used to produce mouse chimeras. The same blot was hybridised sequentially with the two probes (neo and exon3), and the results showed that 14/35 G418 colonies analysed displayed the correct integration pattern on both sides, amounting to

A.



B.

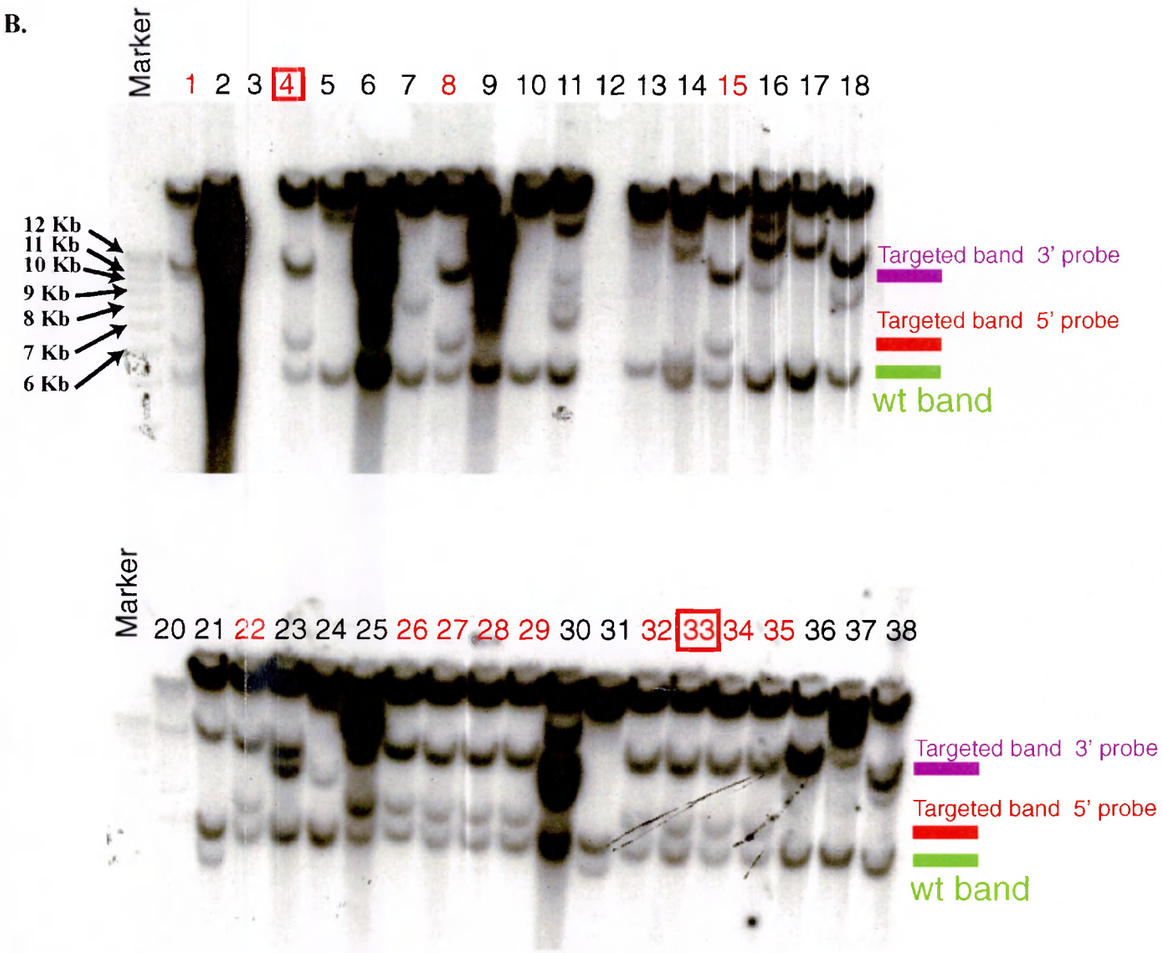


Figure 20 Af4 Targeting in mouse ES cells (Southern hybridization)

A. Graphical representation of the wild type Af4 locus, the Af4-loxP-βgeok targeting construct, and the targeted Af4 allele after homologous recombination in ES cells.
The names are indicated in bold and underlined. The azure arrows on both sides of the wt Af4 locus and the targeted Af4 allele indicate the flanking genomic regions. The 5' probe is represented by a green rectangle within the third exon of the gene. The wt bands identified by the 5' and 3' probes are represented by green lines.
The targeted bands identified by the 5' and 3' probes are represented by a red and a purple line, respectively.
The purple thin boxes represent the 5' and the 3' border of the Af4-loxP-βgeok targeting construct. The positions of the NheI sites are indicated, on which the Southern strategy is based.
The scale (1200) indicates the number of nucleotides/centimeter.
Abbreviations: (sA) splice acceptor; (pA) polyadenylation element; (ECMV IRES) encephalomyiocarditis virus internal ribosomal entry site; (LacZ-neo) fusion of the βgalactosidase and the neomycin genes; (Hygro) hygromycin phosphotransferase gene; (SV40) simian virus 40.

B. Southern blot of NheI digests from the DNA of G418 resistant ES cell clones.
The filters were sequentially hybridized with the 5' and the 3' probe, since both are compatible with an NheI based Southern strategy. 14 out of 35 colonies analyzed showed the correct pattern, amounting to a 40% frequency of homologous recombination. (positive colonies are marked in red). Colonies n. 4 and n.33 (red squares) were used for the generation of chimeric mice.

an extremely high homologous recombination frequency of 40%. This great efficiency is compatible with a promoter trap selection, and clearly argues that *Af4* is expressed in mouse ES cells. In agreement with this prediction, there was good correspondence between the X-gal staining pattern of the G418 selected 48-well plate and the Southern blot results. In fact, most homologously recombined colonies expressed lacZ at approximately the same level (colonies n. 1, 4, 15, 22, 26, 28, 32, 33 and 35 in figure 19) Colonies n. 8, 20, 27, 29 and 34, on the other hand, showed a lower level of B-galactosidase expression. However, it should be noted that X-gal staining is a very qualitative type of assay. Furthermore, there is definitely an element of stochastic variability when one examines the activity of a promoter (in this case the promoter of the *Af4* gene) within the context of the genome packaged into chromatin. For example, the two alleles of a gene might lie, at a given time, in slightly different chromatin environments, so that their respective expression levels might indeed not always be equal. As a consequence, the expression of a selectable marker would vary according to which allele had been targeted. In a population of cells, the weight of stochastic variations in chromatin accessibility and localised availability of appropriate transcription factors is likely to be even greater. It is therefore not surprising to observe, even among the supposedly homogeneous population of G418 resistant correctly targeted clones, a significant degree of variability in the X-gal staining.

On the basis of the above results, clones 4 and 33 were injected into mouse blastocysts (from strain C57Bl/6J) to generate *Af4*LacZ chimeras.

From clone 4, the following chimeras were obtained with level of chimerism, as judged by coat colour, indicated in parentheses:

1 male (100%); 2 males (90%); 3 males (80%); 2 males (60%); 2 males (50%); 1 male (40%); 1 male (30%); 1 female (90%) and 1 female (50%).

The three males with 100% and 90% chimerism were bred to C57Bl/6J females to transmit the allele through the germline.

From clone 33, the following chimeras were obtained:

2 males (100%); 3 males (90%); 2 males (80%); 1 male (70%); 2 males (40%); 1 female (80%); 1 female (70%); 1 female (60%) and 2 females (20%).

Two males with 100% chimerism, one male with 90% chimerism and one male with 80% chimerism were bred to C57Bl/6J females to transmit the allele through the germline.

VIII.4 Transmission of the *Af4*-loxP-βGeok allele through the mouse germline.

In order to transmit the *Af4*-loxP-βGeok allele through the mouse germline, high percentage chimeras from both ES clones 4 and 33 were crossed with wild type mice of the strain C57Bl/6J. Both clones transmitted the allele through the germline, demonstrating that the ES cells had retained their totipotency during the targeting experiment.

Two schemes of breeding were then established to exploit the multiple functions of the *Af4*-loxP-βGeok allele .

F1 mice carrying one copy of the recombined allele were intercrossed to obtain homozygously mutant mice, which could then be used to characterise the degree to which the sA-IRES-βGeok cassette in intron 3 interferes with *Af4* gene activity and the ensuing phenotype.

DNA from mouse tails was analysed by Southern blot hybridisation with the probe ex3 (see above). As shown in figure 21, homozygous mutant mice were born at the expected Mendelian frequency (25%). This shows that the possible impairment caused by the sA-IRES-βGeok cassette did not result in embryonic lethality. This was expected, based on the

results of a previous study, where loss of function of the *Af4* gene was compatible with life (Isnard et al., 2000).

In the second breeding scheme, both *Af4*-loxP-βGeok chimeras and *Af4*-loxP-βGeok heterozygotes were crossed to either one of two Cre expressing mouse lines, to delete the βGeok cassette leaving a single loxP site in the third intron of the gene. The two Cre expressing mouse lines were the X-linked Cre-deleter line (Schwenk et al., 1995) and the PGK-Cre line (Lallemand et al., 1998). Crosses with both lines were successful in Cre mediated deletion of the intervening cassette. Figure 22 shows the results from a representative litter. This proved *in vivo* that a loxP site placed in the *Af4* locus is accessible to Cre mediated recombination, a prerequisite for inducing later the interchromosomal translocation.

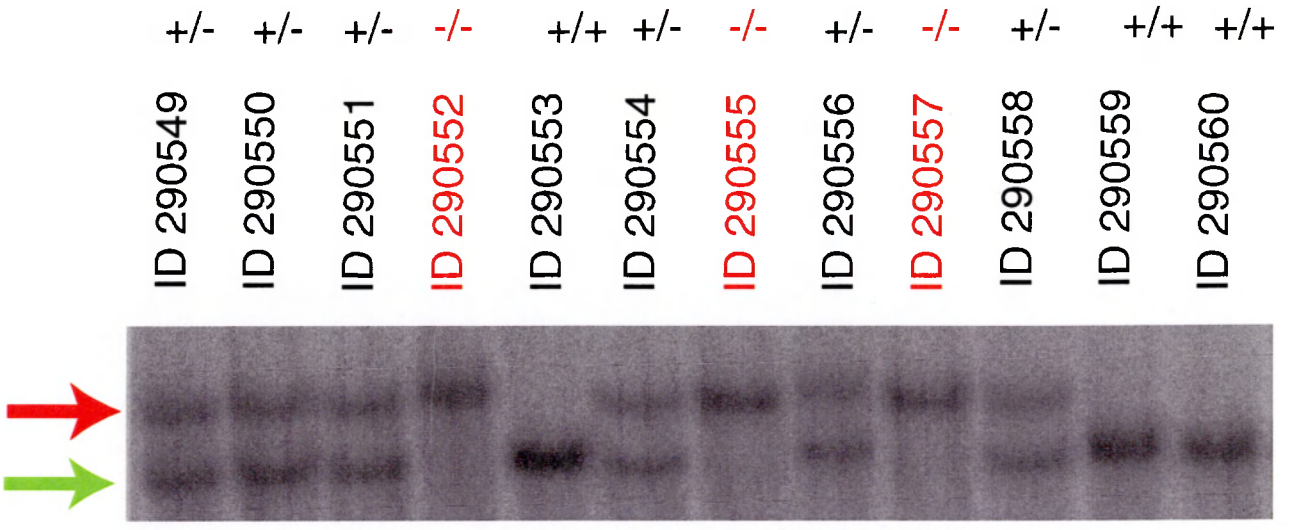
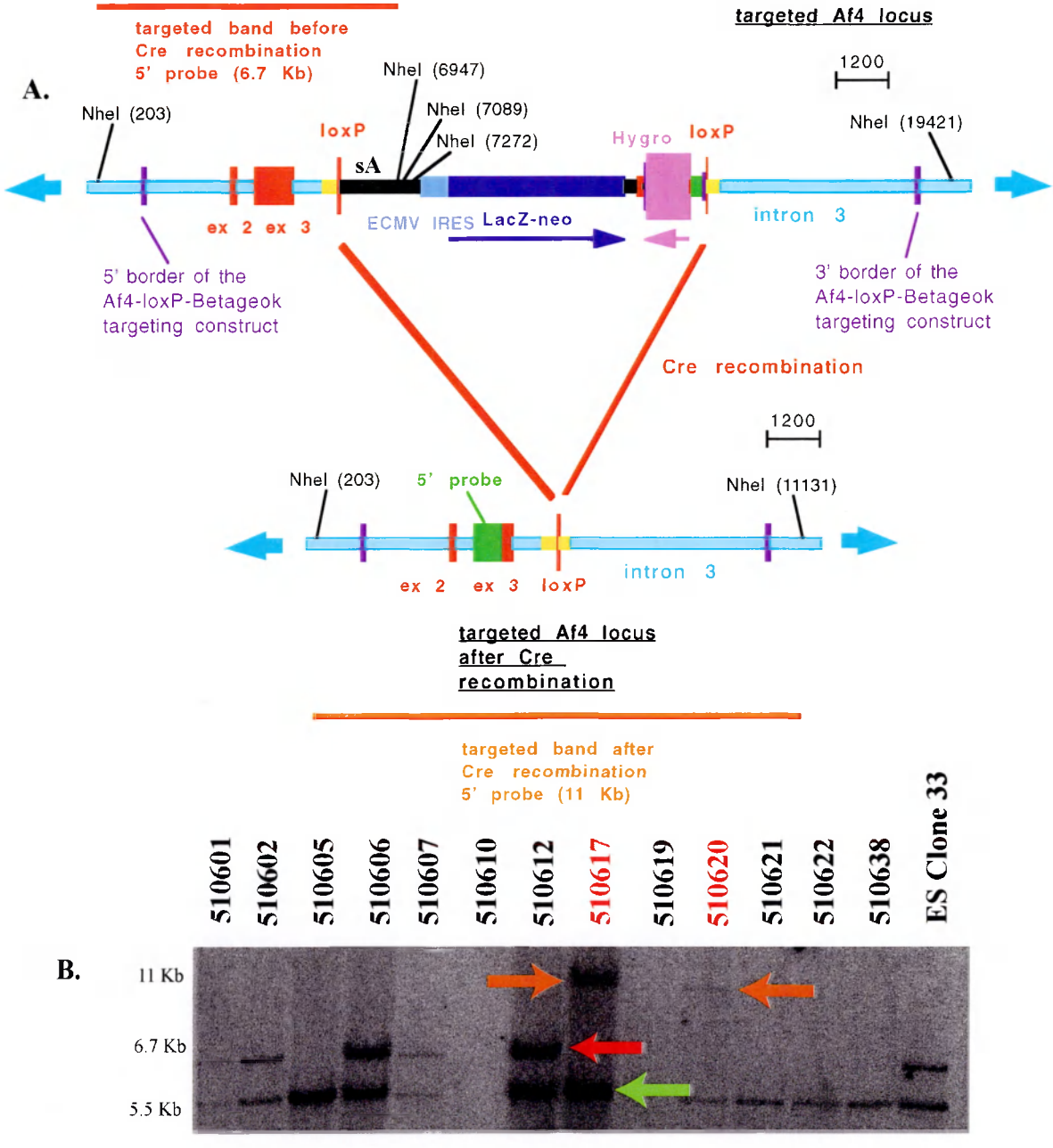


Figure 21 Af4-/- mice are born at the expected Mendelian rate
Southern blot hybridisation on mouse tails from a representative offspring of two heterozygous Af4-loxP-βGeok mice (ID 510639 and 510646). The Probe-ex3 was used, which detects a wild type band of 5546 bps (green arrow) and a recombined band of 6744 bps (red arrow).
Homozygous Af4-loxP-βGeok mice were born at the expected mendelian frequency (3/12=25%). 25% of the mice were wild type, and 50% were heterozygous.



re 22 Cre deletion of the mutant Af4 allele

raphical representation of the targeted Af4 locus before and after Cre recombination.
ames are indicated in bold and underlined. The azure arrows on both sides of the Af4 locus indicate
lanking genomic regions. The 5' probe (ex3) is represented by a green rectangle within the third exon
gene. The targeted band before Cre recombination identified by the 5' probe is represented by a red line.
argeted band after Cre recombination identified by the 5' probe is represented by an orange line.
urple thin boxes represent the 5' and the 3' border of the Af4-loxP-βgeok targeting construct.
ositions of the NheI sites are indicated, on which the Southern strategy is based.
cale (1200) indicates the number of nucleotides/centimeter.
eviations: (sA) splice acceptor; (pA) polyadenylation element; (ECMV IRES) encephalomyiocarditis
internal ribosomal entry site; (LacZ-neo) fusion of the βgalactosidase and the neomycin genes; (Hygro)
mycin phosphotransferase gene; (SV40) simian virus 40.
presentative Southern hybridization analysis of a litter from an Af4-loxP-βgeok X Cre-deleter mouse
ross. Analogous results were obtained with the PGK-Cre mouse line. The 5' probe (ex3) was used to
t the band (orange arrow) resulting from Cre deletion of the loxP-flanked βGeok-hygro cassette on an
digest of mouse tail genomic DNA. The red arrow points to the targeted band, prior to Cre recombination,
ted with the same exon 3 probe. The green arrow points to the wild type band, detected with the same exon
be.

VIII.4.1 Mice homozygous for the *Af4*-loxP-βGeok allele do not express the full *Af4* transcript

To determine whether the loxP flanked βGeok cassette in intron 3 does actually interfere with transcription of the *Af4* gene, an RT-PCR strategy was designed. RNA was extracted from both wild type and *Af4*-loxP-βGeok homozygous mice, and cDNA was synthethised. Three primers were used to explore different regions of the *Af4* transcript, together with two primers to amplify βactin as a positive control for the quality of the RNA and the efficiency of the RT-PCR conditions used. Primer *Af4*ex3F and primer *Af4*ex11R amplify the *Af4* transcript from exon 3 to exon 11, yielding a band of 1430 nucleotides. Primer *Af4*ex3F has the following 5'-3' sequence:

CTTTATCGATTGGGTGACTATGAGGAGATGAA

Primer *Af4*ex11R has the following 5'-3' sequence:

CTGCCCTCAGCGACACCCTTACTT

The β-actin primers have the following 5'-3' sequences:

β-actinF (GGCCCAGAGCAAGAGAGGTATCC)

β-actinR (ACGCACGATTTCCCTCTCAGC). They yield a band of 439 nucleotides. Since the βGeok cassette is placed in intron 3, the *Af4*ex3F-*af4*ex11R primer pair allows to check whether this configuration prevents full transcription of the *Af4* gene. As expected, no transcript is detected in the *Af4*-loxP-βGeok homozygous animals, while the β-actin band is readily detected in both wild type and homozygous mice (figure 23). In case that some nascent transcripts would escape processing by the SV40 late polyadenylation signal of the knock-out cassette, aberrant mRNAs could be produced by cryptic donor sites in the βGeok cassette splicing to exon 4. Due to their length, these transcripts would most probably not be

detected by the RT-PCR assay described above. However, even if such mRNAs were produced, they would still result in a prematurely truncated protein after exon 3.

To check whether the βGeok cassette also results in the absence of any downstream *Af4* transcript, the same primer *Af4*ex11R was used in combination with primer *Af4*ex4F, which anneals in the fourth exon of *Af4*, and has the following 5'-3' sequence:

GAAGTCCATTGCGTTGAAGAGATT

Surprisingly, a band of the expected size was detected in both wild type and *Af4*-βGeok homozygous animals. A suggestive explanation for this result is the presence of an alternative promoter located downstream of the βGeok cassette. At present, it is not known whether in homozygous mice this 5' terminal truncated transcript is actually translated in the corresponding aminoterminally truncated protein. Even if this were the case, these animals would still lack the full AF4 protein lacking both the NHD (N-terminal homology domain) and the first half of the ALF domain, as defined in (Nilson et al., 1997). As discussed in chapter 6, both these domains are highly conserved between all members of the *AF4* family, suggesting that such a truncated protein would be very likely impaired in its function. Furthermore, it could also behave as a dominant negative.

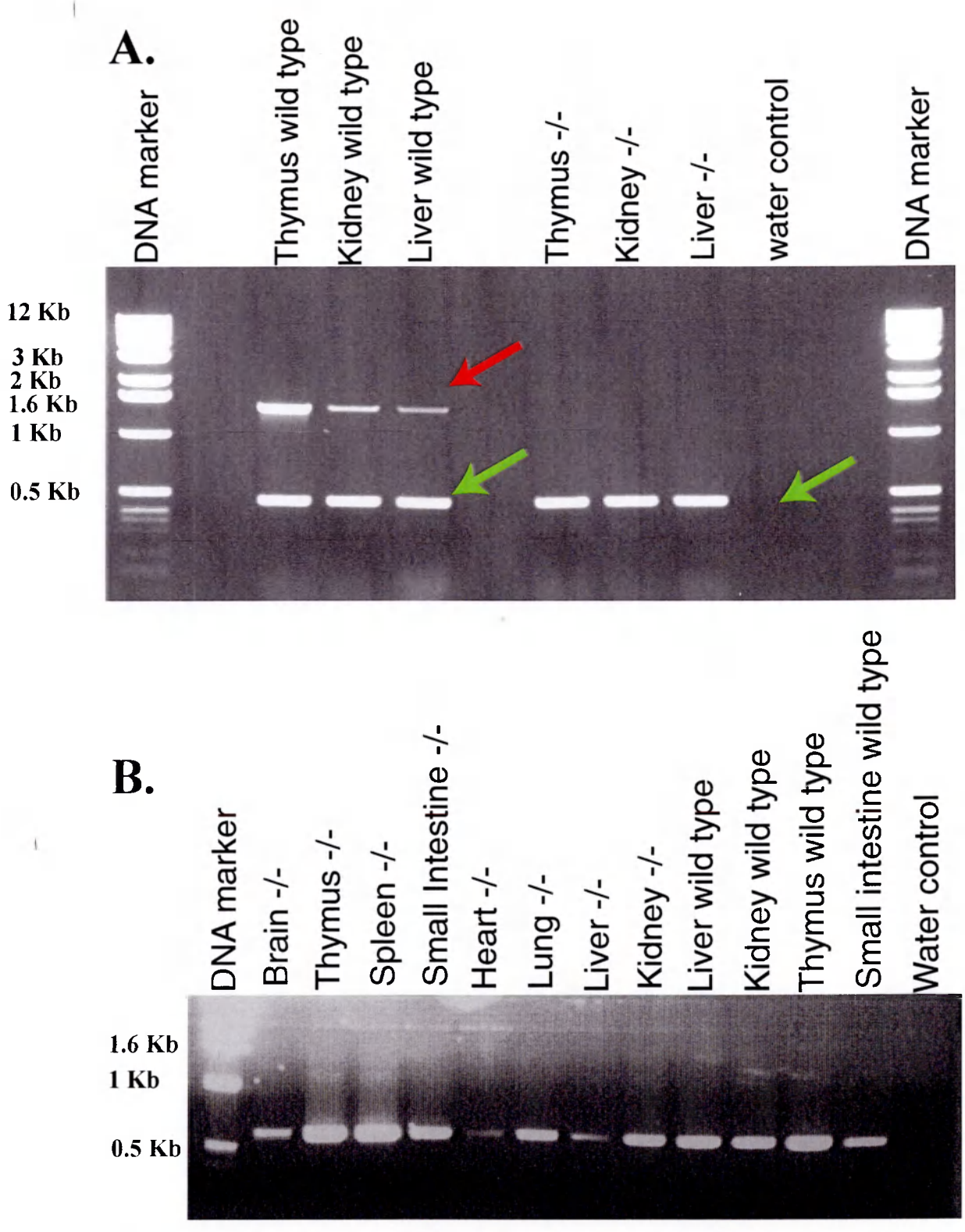


Figure 23 RT-PCR analysis of *Af4*-loxP-βGeok homozygous mice
An RT-PCR analysis was performed on several organs from mice homozygous for the *Af4*-loxP-βGeok allele.
A. RT-PCR with primers Af4ex3F, Af4ex11R, βactinF and βactinR. Primers Af4ex3F and Af4ex11R amplify the Af4 transcript (red arrow) only in wild type mice, while the βactin band (green arrow) is readily detectable in both wild type and homozygous mice. This indicates that no full transcript is produced in the *Af4*-loxP-βGeok homozygous animals.
B. RT-PCR with primers Af4ex4F and Af4ex11R. This experiment explores the presence of the *Af4* transcript downstream of the βGeok cassette inserted in the third intron of *Af4*. The presence of the band in both wild type and *Af4*-βGeok homozygous animals argues for the presence of an alternative promoter located most likely in intron 3 downstream of the knock-in cassette.

IX

Engineering of a multifunctional *Mll* allele

IX.1 Overview of the strategy

Gene targeting in the mouse has been extensively employed to analyse *in vivo* the function of many genes through the generation of knock-out, knock-in and hypomorphic alleles. As the technique has been progressively streamlined, it is becoming increasingly possible to address questions in the context of the whole organism.

One of the main assets of mutational analysis is the ability to assess the role that single domains, sometimes even single residues, play in the function of genes and proteins *in vivo*. In the context of modern mouse genetics, this is usually achieved by generating different targeting vectors for any desired mutation, a very time consuming process. Indeed, besides the logistics of suitable animal facilities, one of the main impediments to generating an allelic series of desired mutations in a gene of interest is the time involved in the assembly of the relevant targeting vectors. Furthermore, the option to combine two or more mutations in cis for assessment of the combination is currently very limited For large genes, where the sites to be mutated are too far apart to be included in the same targeting vector, again individual targeting constructs need to be generated for each desired mutation. To target two mutagenic cassettes to the same locus, mouse ES cells must be subjected to several rounds of electroporation and drug selection, a procedure which can result in loss of totipotency and subsequent inability to colonise the mouse germline. Additionally, upon retargeting a given locus at a new position with a second cassette, only 50% of the homologously recombined clones will have the two targeting cassettes on the same allele, and only these will be useful for propagating the desired mutations together through the mouse germline. Particularly for genes which display low homologous recombination

efficiencies, the impact of a 50% reduction in the amount of successfully targeted ES clones cannot be overemphasised.

Furthermore dual mutational combinations need not be only changes in amino acid or cis element sequence. Introduction of a reporter gene (like β-galactosidase or GFP) or of a protein purification tag are both examples of useful mutations for gene function analysis *in vivo*. Targeting of many genes with GFP and/or β-galactosidase is now a widespread approach to characterising, at the organismal level, expression patterns throughout development and intracellular localisations. To apply these tools to the study of the effect of specific mutations on expression patterns and/or the intracellular localisations, the fluent introduction of combinations on the same allele of a deleterious mutation and a reporter gene, presents new options for mutational analyses.

This approach can also encompass biochemical analysis using protein-purification tags fused in frame to the protein of interest. Systematic characterisation of biochemical complexes is an ambitious goal of modern large scale biology, commonly referred to as proteomics, and a prerequisite to understanding physiological and pathological processes in molecular detail. A variety of protein tags have been applied to characterise multicomponent complexes using reliable, generic purification protocols. The use of protein tags in knock-in experiments presents another use of directed mutagenesis in the mouse and a further example for combinatorial two-site mutagenic applications.

Ideally, for any given gene, a full battery of targeting constructs can be envisioned, each combining a protein-tag fusion to facilitate *in vivo* localisation studies or purification of protein complexes with the wt or specific mutations which explore the relevance of different domains or residues.

It is evident that realisation of the potential inherent in dually mutagenised alleles lies with the ability to rapidly establish ES cell lines harbouring the combinatorial allele for the

gene of interest. The strategy outlined in this chapter for the *Mll* gene presents a novel approach to a dual engineering task and an experimental platform to test its feasibility.

IX.2 Establishment of an *Mll* allele mutated at two sites 60 kb apart

The primary goal was to insert a loxP site in the *Mll* gene at an equivalent position to the BCR for translocation with the *Af4*-loxP allele. For many of the same reasons discussed above for *Af4*, I chose not to position only a loxP site but also to create a multi-purpose allele. Hence, the same βGeok cassette (sA-IRES-βGal-neo-pA) was used after placement between loxP sites. As for *Af4*, the *Mll*-βGeok allele serves two purposes before Cre recombination. First, β-galactosidase activity can be used to report *Mll* promoter activity in situ. Second, the SV40 polyadenylation signal should truncate the primary transcript to produce, possibly, an interesting hypomorphic allele. If expressed and stable, the truncated MLL protein will correspond to the leukemogenic MLL part of MLL fusion proteins without a fusion partner. However, in contrast to the *Af4* work, the possibilities presented by the *Mll*-βGeok allele are partially redundant to previous work on *Mll* (Dobson et al., 2000; Yagi et al., 1998). Nevertheless, we felt that inclusion of the βGeok strategy here had sufficient merit. Also, other work in the Stewart lab had shown that βGeok was the most reliable cassette at hand for the primary goal anyway.

The secondary goal in creation of this *Mll* allele was the addition of a protein tag (TAP-tag see below) to the very N-terminus. Thereby I hope to utilise the protein tag for biochemical purifications and/or immunoprecipitations of MLL in various ways. They are summarised here and illustrated in figure 24.

1) Before Flp recombination, the TAP-tag is not fused to MLL and serves as a good control for biochemistry. This allele (TAP-*Mll*-hygro-IRES-LacZ) is predicted to result in a complete loss of function. The hygromycin cassette stops translation of the protein after the end of the ATG-TAP-FRT fusion. Even assuming the presence of a downstream promoter, the trap cassette in intron 11 should be sufficient to impede generation of a full length

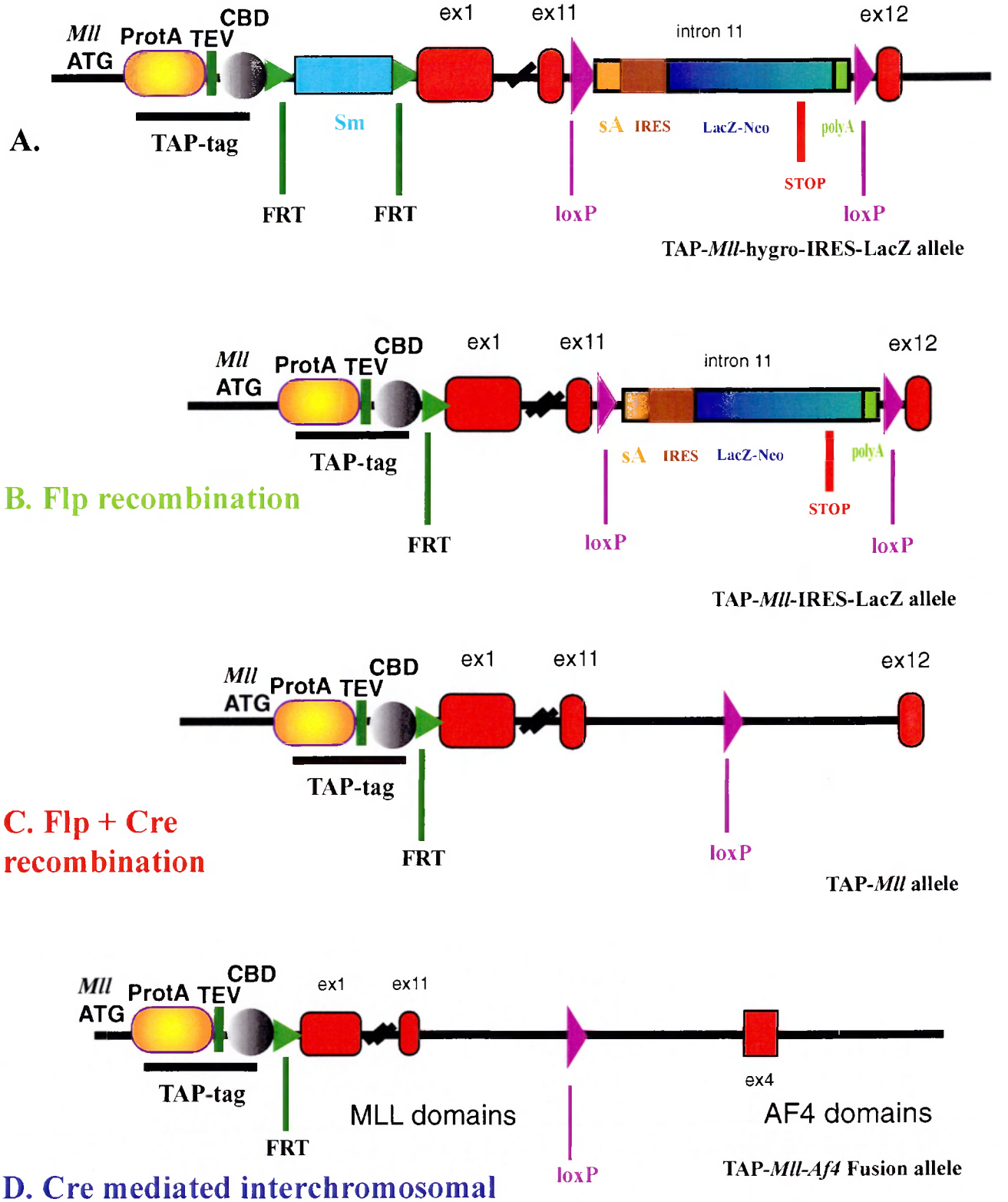


Figure 24 Four *Mll* alleles generated from a single targeting construct
Schematic representation of the four alleles which can be generated from the single *Mll* targeting construct using Cre and Flp recombinases. The allele depicted in panel A should ablate expression of the gene. The alleles in panels B, C and D are expected to allow biochemical characterisation of the truncated, wild-type and translocated MLL proteins, respectively.
See text for a full description of the potential uses of these alleles.
Abbreviations: CBD (calmodulin binding domain); sA (splice acceptor element); IRES (internal ribosomal entry site); polyA (polyadenylation signal); TAP (tandem affinity purification)

transcript. In fact, in a previous study, deletion of exons 12 through 14 resulted in a complete loss of function (Yagi et al., 1998).

2) After Flp recombination and before Cre recombination, the TAP-tag should be fused to the N-terminal half of MLL without the C-terminal half (TAP-*Mll*-IRES-LacZ allele). This should in principle allow characterisation of the protein complex assembled around the N-terminal domains of MLL from a variety of tissues and organs. This could yield precious insights into MLL leukemogenic potential, as the truncated protein contains the domains which are consistently retained in the MLL fusion proteins (the three AT hooks, the subnuclear localisation signals and the CXXC domain). Furthermore, with this allele Lac-Z staining could be used to analyse the expression pattern of *Mll* throughout embryonal development and adult life.

3) After Flp and Cre recombination, the TAP-tag is fused to full length MLL (TAP-*Mll* allele). This should allow characterisation of the wild-type protein complex(es) in which the MLL protein exerts its function. Comparison of this protein complex with the TAP-*Mll*-IRES-LacZ complex could help delineate which protein interactions are centered around the N-terminal (leukemogenic) half of the protein as opposed to the C-terminal.

4) After Flp and Cre recombination, followed by Cre mediated translocation to *Af4*, the TAP-tag is fused to the MLL-AF4 leukemogenic protein. If leukemias arise, it may be possible to obtain these cells in sufficient quantities for biochemistry. Furthermore, this mouse line could also be used to address one of the most puzzling issues regarding MLL oncogenic potential. In spite of widespread *MLL* expression throughout development and adult life, so far *MLL* translocations have been observed only in leukemias. As discussed previously, there could be several reasons for this, which can be grouped into either a pre-recombinational or a post-recombinational bias. Driving interchromosomal translocations in a variety of compartments other than the blood lineages could be very informative. If no

other tumors arise in spite of the translocation, western blotting could be used to probe the translocation positive cells for the presence of partner proteins identified in the leukemogenic complex. Their possible absence would help to understand why in these cases the fusion protein has a neutral function.

In cases (2), (3) and (4), the tagged proteins will be heterozygotic to the wt protein, potentially permitting to address questions of multimerisation in protein complexes. In case (3), homozygosity of the tagged protein can be used to evaluate the phenotypic impact of the tag. Furthermore, the TAP-tagged alleles should facilitate biochemical investigations during development and in different tissues. In particular, it will be interesting to compare associated proteins in ES cells, representing the totipotent, stem cell condition, to differentiated tissues.

Finally, the various configurations of the *Mll* allele potentially create an allelic series of three from knock-out to neutral. Together with heterozygotic combinations of the three configurations, a spectrum of graded phenotypic severity may also be obtained.

IX.2.1 The tandem affinity purification (TAP) system

The tandem affinity purification (TAP) protein tag was chosen because biochemical purification of proteins expressed at their native levels poses a greater challenge than approaches relying on overexpression (Rigaut et al., 1999). The TAP was shown in yeast to enable accurate purification of protein complexes present at relatively low levels. It consists of a fusion of the IgG-binding unit of protein A of *Staphylococcus aureus* (ProtA) with the calmodulin binding domain (CBD). A TEV protease cleavage site is present between the two moieties. The first affinity purification step utilises IgG coated beads which bind to the ProtA element. Following TEV cleavage to free the protein complex, a second round of purification is carried out using calmodulin coated beads. Importantly, all steps enable

specific release from affinity columns under native conditions, one of the strongest aspects of this system.

IX.2.2 Construction of the doubly mutated *Mll* allele

In order to select for integration into the *Mll* locus, both in an E.Coli BAC and in the ES cell genome, the TAP module was linked to a hygromycin resistance cassette, flanked by FRT sites, so that upon Flp mediated recombination, the hygromycin resistance gene is deleted. In this configuration, after the first methionine coded by the starting ATG of the *Mll* gene follows translation of the TAP tag in frame with the FRT left after Flp recombination; immediately downstream of the FRT, translation of the MLL protein, from the second amino-acid onward, resumes.

The second mutagenic cassette is a loxP-flanked sA-IRES-βGeok-pA module inserted in the intron 11 of the *Mll* gene. This intron is one of the most frequently disrupted in *Mll* translocations. As already discussed, insertion of this cassette in an intron has been shown to result in truncation of the normal mRNA of the gene (which may or may not be translated) and expression of the βGeok protein under the regulatory control of the targeted gene. Upon Cre mediated recombination, this whole cassette is deleted, leaving one single loxP site in intron 11. Importantly, this can be then used to drive interchromosomal translocation with the *Af4*loxP mouse line.

The experimental steps involved in generating this targeting construct are summarised below:

1. Isolation of a mouse *Mll* BAC.
2. Characterisation of the relevant regions of the BAC through:
    2.1 ET mediated isolation of the promoter region
    2.2 Direct BAC sequencing

3. Assembly of the *Mll* ES targeting construct by ET recombination:

3.1 ET mediated engineering of the BAC-based backbone for the targeting construct.

3.1.1. ET mediated deletion of the 5' side of the *Mll* BAC

3.1.2. ET mediated deletion of the 3' side of the *Mll* BAC

3.2 Assembly of the knock-in cassette *Mll*Hom5-TAP-FRT-pA-hygro-PGK-FRT-*Mll*Hom3

3.3 Assembly of the knock-in cassette int11-loxP-sA-IRES-βGeok-pA-loxP-int11

3.4 ET mediated insertion of the int11-loxP-sA-IRES-βGeok-pA-loxP-int11 cassette into the BAC-based construct backbone.

3.5 ET mediated insertion of the *Mll*Hom5-TAP-FRT-pA-hygro-PGK-FRT-*Mll*Hom3 cassette into the BAC-based construct backbone.

IX.3 Assembly of the *Mll* targeting construct

IX.3.1 Isolation of the mouse *Mll* genomic clone from a high density BAC library

High density mouse BAC filters (Research Genetics, Inc.) were screened with a mouse *Mll* specific probe (pS2). The probe was a cDNA fragment corresponding to the first three PHD fingers of the MLL protein (PH-A, PHD-B and PHD-C) which span exons 11 through 16.

Hybridisation of this probe to the high density BAC filters yielded 5 independent positive signals (Figure 25). 3 clones, n.146D23, 145K16 and 145L16, screened by PCR and Southern hybridisation, were found to contain bands of the expected size. (Figure 26)

A combination of both direct sequencing, PCR amplification and Southern blot hybridisation was used to check for the presence, within the same BAC, of both the first exon of the gene (as a substrate for targeting the TAP cassette) and intron 11 (as a substrate for targeting the loxP-βGeok cassette).

The probe used for Southern hybridisation was an oligonucleotide probe (m*Mll*ex1) with the following 5'-3' sequence:

GGCCGGCCCTGCTCCGGGTGGGCCCGGGCTTCGACGCGGCGCTGCAGGTCTCGG

CCGCCATCGGCACCAACCTGCGCCGGTTCCGGGCAGTGTTTGGGGAGAGCGGCG

GGGGAGGCGGCAGCGGAGAG

BAC 145L16 was found to contain both regions of the *Mll* gene needed for targeting and was therefore selected for the further steps.

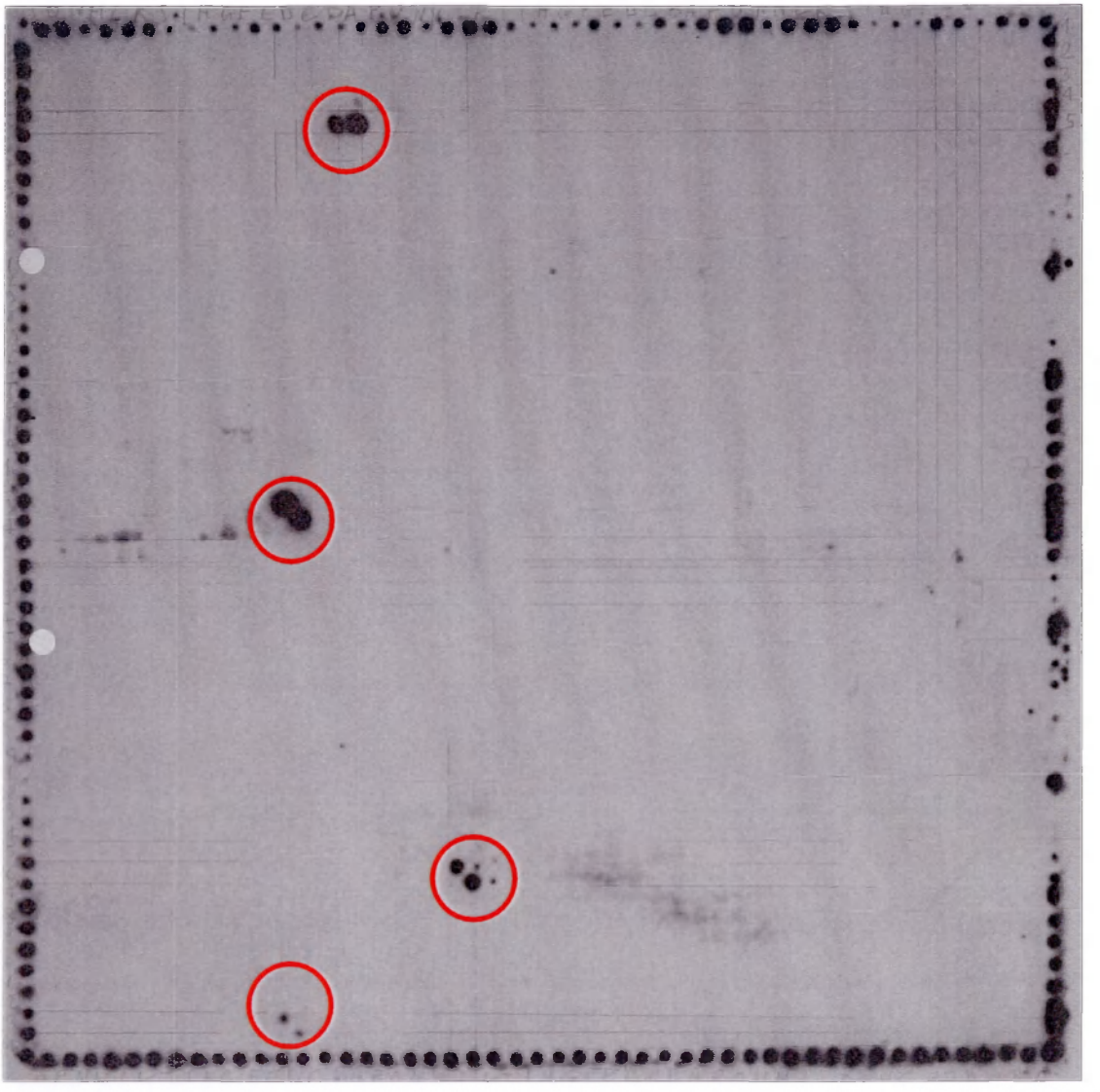


Figure 25 Isolation of Mll BAC clones
Hybridisation of high density mouse BAC membranes with an Mll probe (pS2) spanning the exons coding for the three PHD fingers (PHD-A, PHD-B and PHD-C). Each BAC clone is spotted twice to distinguish true signals (a double spot) from false positives.The filter was simultaneously hybridised with a control probe which highlights the reference marks on the four sides of the membrane to unequivocally identify the clone of interest.
Positive clones are circled in red.

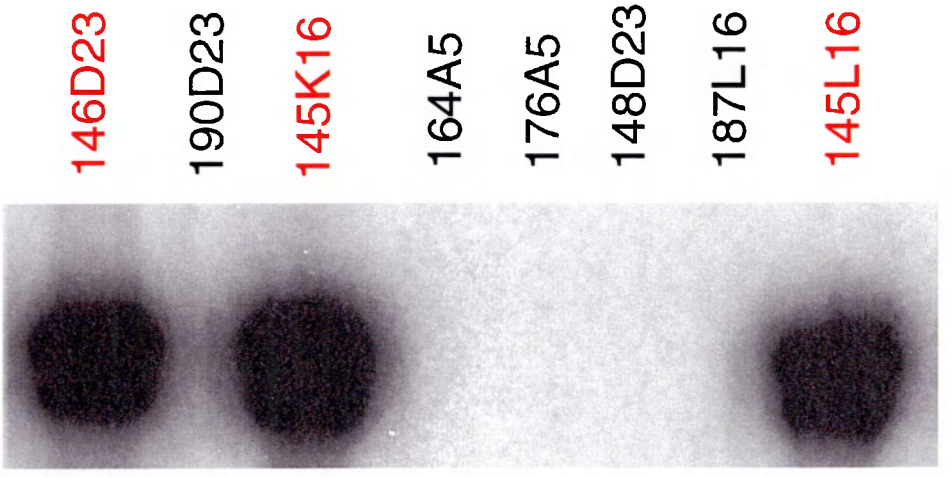


Figure 26 Southern blot hybridisation of candidate Mll BACs
8 candidate Mll BACs isolated from a BAC library screening were tested with the Mll probe pS2. BAC clones n. 146D23, 145K16 and 145L16 display the correct signal. BAC 145L16 was chosen for further characterisation.

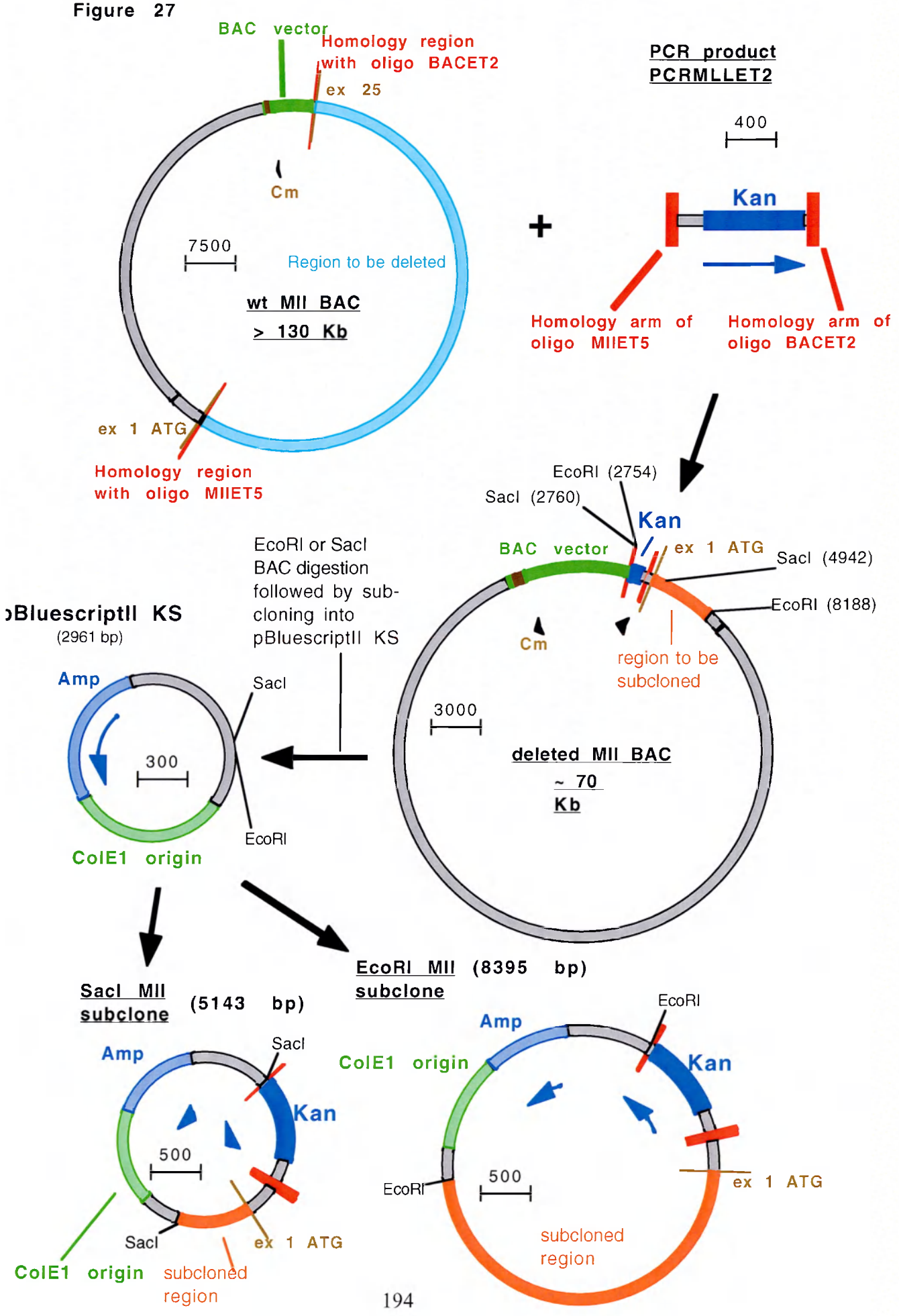
IX.3.2 ET mediated cloning of the 5' region upstream of *Mll* exon 1

Once both sequence analysis and Southern blot hybridisation had confirmed the presence of *Mll* exon 1 in the BAC 145L16, it became important to derive more sequence information on the region immediately upstream of the initiating ATG of *Mll* exon 1 so that a Southern blot strategy which would discriminate between random integration and homologous recombination events could be developed. Also, the substantial length of extra 5' homology present in the BAC would make it impossible to devise a suitable Southern strategy and hence needed to be eliminated. Therefore, sequence information was also needed to devise a strategy which would shave the BAC from one end of the BAC vector up to about 5 kb upstream of the initiating ATG. In this step a rare cutting restriction enzyme site was concomitantly introduced so that the final targeting construct could be easily severed from the BAC vector prior to ES cell electroporation. In order to perform this ET cloning mediated deletion ("BAC shaving"), it was necessary to obtain sequence information from the region around 5 kb upstream of the ATG.

While the 5' region of the human *MLL* gene has recently been sequenced, the corresponding region in the mouse was completely unknown. A primer-walking sequencing approach would have been very time consuming. Plus, for BACs to become the tools of choice in contemporary genome engineering, techniques must be developed which improve their ease of manipulation, including the ability to rapidly isolate a region of choice from a BAC. Therefore, I decided to apply a further variation of the ET methodology to subclone this region from the BAC. In this variation (figure 27) a region of choice is deleted from the target DNA molecule and replaced with a selectable marker. Here, all of the *Mll* BAC downstream of exon 1 to the BAC vector was deleted. This placed a selectable marker immediately adjacent to the unknown region to be subcloned on one side, and the known sequence region of the BAC vector on the other. Upon digestion of this deleted BAC with

Figure 27 ET mediated BAC deletion to subclone the promoter region of Mll
An ET recombination approach was chosen to delete from the Mll BAC the whole region spanning exon 1 to exon 25, by replacing it with the kanamycin resistance gene. The Mll promoter region was then easily isolated by random BAC subcloning into the vector pBluescriptII KS applying kanamycin selection.
The names of the DNA molecules (constructs and PCR products) are in bold and underlined. The scale bar indicates the number of nucleotides/centimeter. The region to be deleted is shown in light blue. The subcloned region is shown in orange. The homology arms and target regions mediating ET recombination are shown in red. Segments of the constructs accompanied by arrows indicate open reading frames
Abbreviations: (Cm) chloramphenicol resistance gene; (Amp) βlactamase gene;
See text for a full description of the cloning strategy.

Figure 27


BAC vector
Homology region with oligo BACET2
ex 25
Cm
7500
Region to be deleted
wt MII BAC
> 130 Kb
ex 1 ATG
Homology region with oligo MIIET5
PCR product
PCRMLLET2
400
Kan
Homology arm of oligo MIIET5
Homology arm of oligo BACET2
EcoRI (2754)
SacI (2760)
Kan
BAC vector
ex 1 ATG
SacI (4942)
EcoRI (8188)
Cm
region to be subcloned
3000
deleted MII BAC
~ 70 Kb
EcoRI or SacI BAC digestion followed by sub-cloning into pBluescriptII KS
pBluescriptII KS
(2961 bp)
Amp
SacI
300
EcoRI
ColE1 origin
SacI MII subclone (5143 bp)
EcoRI MII (8395 bp) subclone
Amp
SacI
Kan
500
SacI
ex 1 ATG
ColE1 origin
subcloned region
EcoRI
Amp
ColE1 origin
Kan
500
EcoRI
ex 1 ATG
subcloned region

appropriate enzymes chosen for suitability from the known sequences and ligation to a standard plasmid, the positive colonies harbouring the region of interest could be very easily identified by doubly selecting for both the plasmid marker and the selectable marker positioned in the BAC.

End sequencing from the BAC vector revealed that up to intron 26 of *Mll* was present and also established the orientation. With this information, the following oligos were designed: *Mll*ET5 and BACET2.

Oligo *Mll*ET5 has the following sequence 5'-3':

TGCTCCGGGTGGGCCCGGGCTTCGACGCGGCGCTGCAGGTCTCGGCCGCC ATCGGCACCAACCTGCGCCGGTTCCGGGCAGTGT*TTAATTAA*AGCAGGTAGCT TGCAGTGGGCTTACATGGCG. Residues 1-86 (in bold) constitute the *Mll* homology arm starting from near the 5' edge of known sequence, 301 base pairs upstream from the initiating ATG of *Mll*. The PCR primer part of this oligo (residues 93-123, underlined) was used to amplify the kanamycin resistance selectable marker. A PacI site was introduced in the oligo as a possible anchor point in subsequent cloning steps (positions 85-92, italic).

Oligo BACET2 has the following 5'-3' sequence:

TCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAATTGTAATACGACTCA CTATAGGGCGAATTCGAGCTCGGTACCCGGGGA*TTTAATTAA*TCGAACCCCAG AGTCCCGCTCAGAAGAACTCG.

Residues 1-85 (in bold) constitute the homology arm to the BAC vector. The PCR primer part of this oligo includes residues 94 to 125 (underlined). Also in this case, a PacI site was introduced in the oligo as a possible anchor point in subsequent cloning steps (positions 86-93, italic). The resulting PCR product (PCR*MLL*ET2) is 1205 base pairs long.

For the ET cloning reaction, HS996 cells, harbouring the *Mll* BAC, were transformed with the recombinogenic plasmid pBAD-αβγ, and made competent as described (see

materials and methods). After electroporation with the PCR product, cells were plated under double selection (chloramphenicol 25μg/ml and kanamycin 25 μg/ml).

Among all colonies analysed, two closely related restriction patterns were observed. Both showed substantial and almost identical deletions. To analyse in more detail the two types of deletions, colonies G and T were chosen for further analysis. Maxiprep DNA preparations were made of both colonies and digested with either EcoRI or BamHI. Figure 28 shows the analysis of these digests by electrophoresis. By comparing the patterns of lanes 1 and 2 (EcoRI digests of wt BAC 145L16 and the deleted BAC-G, respectively), and the patterns of lanes 3 and 4 (BamHI digests of the wt BAC 145L16 and the deleted BAC-G respectively), it was possible to estimate the extent of the deletion as being greater than 50 kb. By comparing lanes 2 with 7 (where BAC-G in lane 2 shows a higher band than BAC-T in lane 7), and lanes 4 with 8 (where BAC-G in lane 4 shows a higher band than BAC-T in lane 8), it also became clear that BAC-G was longer than BAC-T by approximately 3 kb (Figure 28).

To further characterise the deletion, pulsed-field-gel electrophoresis was performed on BamHI digests of the wt BAC and the deleted BAC-G. (Figure 29)

Finally, to subclone and characterise the region 5' of the ATG, DNA preparations of both colonies G and T were digested with either EcoRI or SacI. These two enzymes were chosen because they only cut the known DNA in the BAC vector near the inserted kanamycin resistance cassette. After ligation to pBluescriptII and analysis, it was found that subclones from both G and T deleted BACs were the same. The reason for the occurrence of two slightly different patterns of deletion was therefore not further investigated, since it did not constitute the main objective of this experiment.

Final sequencing of EcoRI and SacI subclones delivered approximately 5.5 kb of unknown upstream sequence.

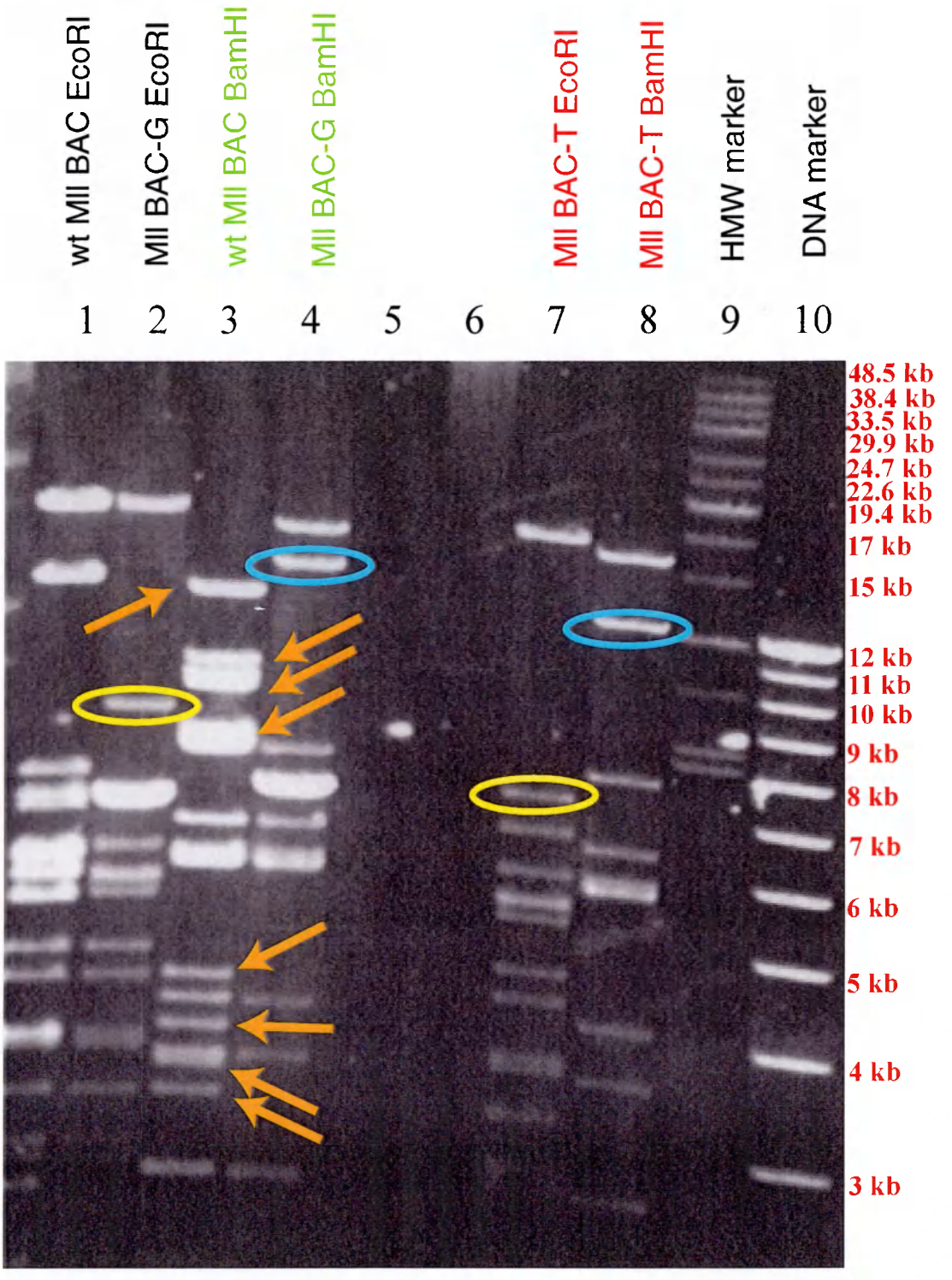


Figure 28 ET mediated BAC deletion to subclone the promoter region of Mll (restriction analysis)

The region from exon 1 to intron 26 was replaced with the kanamycin resistance gene by ET recombination. The extent of the deletion was estimated by electrophoresis of EcoRI and BamHI digests. Orange arrows indicate the bands of the wt Mll BAC which are missing from two independent "ET deleted" colonies, Mll BAC-G and Mll BAC-T. More than 50 kb were deleted.

Mll BAC-G and Mll-BAC-T differ only by about 3 Kb, as indicated by the bands circled in yellow and azure from EcoRI and BamHI digests respectively.

HMW= high molecular weight DNA marker (Gibco, BRL)

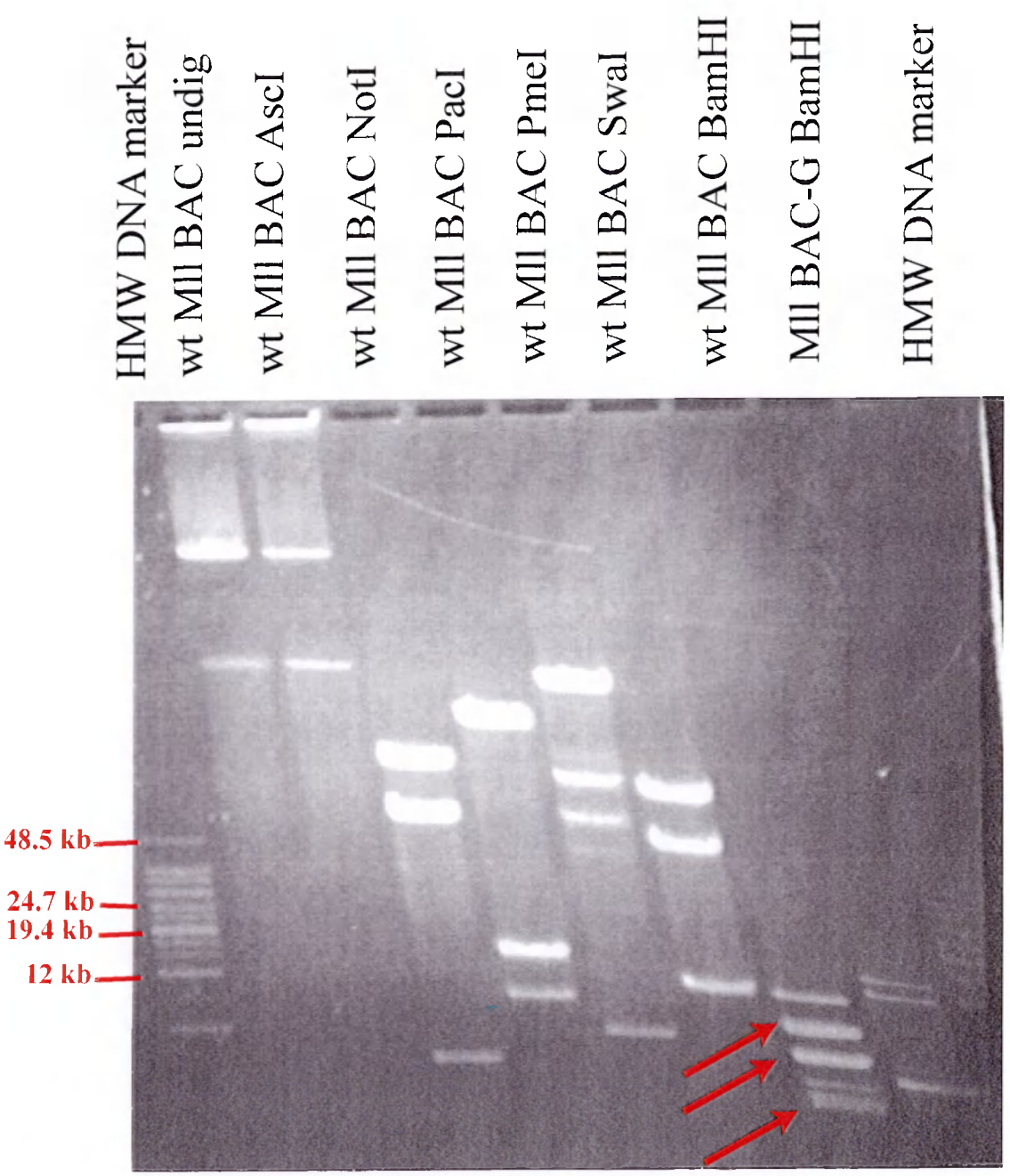


Figure 29 Pulsed field gel electrophoresis of wt and "ET deleted" Mll BACs
The wt Mll BAC, as well the "ET deleted" Mll BAC-G were analysed by pulsed field gel electrophoresis. Red arrows indicat the bands of the wt Mll BAC which missing from the deleted Mll BAC-G

IX.3.3 ET mediated engineering of the *Mll* BAC to yield the backbone for the *Mll* targeting vector.

Having obtained sufficient sequence information for the regions 5' of the initiating ATG and 3' of intron 11, a strategy to shave the *Mll* BAC on both sides (figure 30) to yield the backbone for the final targeting construct was undertaken.

HS996 cells harboring the wt *Mll* BAC 145L16 were transformed with the recombinogenic plasmid R6Kαβγ. Two colonies, n. 3 and n. 8, were used to prepare batches of competent cells.

In the first step of the deletion strategy, a PCR product was designed to replace the 5' end of the *Mll* BAC from the BAC vector to approximately 5 kb upstream of the initiating ATG. To this end oligos *Mll*Del1F and *Mll* Del1R were used to amplify the β-lactamase gene.

Primer *Mll*Del1F has the following 5'-3' sequence:

TTCACACAGGAAACAGCTATGACCATGATTACGCCAAGCTATTTAGGTGACACT

ATAGAATACTTAATTAATGAAGACGAAAGGGCCTCGTGATACGCC.

Residues 1-63 constitute the arm of homology to the BAC backbone. A PacI site was included at positions 64-71 to facilitate subsequent recognition of correctly targeted BACs. The portion of this oligo annealing to the β-lactamase gene (from plasmid pACYC177) comprises residues 72 to 99.

Primer *Mll*Del1R has the following 5'-3' sequence:

GCAGGTACCTAGCCATATGCCTGTTTCCTCATTTGCAAACATAAGAATATTAATA

GCAACTCCTGCCATTTCATTACCTCTTTCTCCGCACCCGACATAGATATCAATCT

AAAGTATATATGAGTAAACTTG.

Residues 1-63 constitute the arm of homology to the *Mll* region 5031 bp upstream from the initiating ATG of the gene. Residues 64-102 constitute a site for the extremely rare cutting

Figure 30 Sequential "shaving" of the Mll BAC

Schematic representation of the sequential "BAC shaving" strategy to obtain a suitable BAC based backbone for the insertion of knock-out or knock-in cassettes. The case of the Mll BAC is illustrated here. ET recombination was used to replace the irrelevant portions of the BAC with two selectable markers. Rare restriction sites (in this case Pi-SceI) were introduced into the ET homology arms, to enable release of the targeting construct from the BAC vector prior to ES cell transfection.

The region to be deleted in the first round is shown in light blue. The selectable marker used in the first round was the β-lactamase gene (Amp), also shown in light blue. The homology arms and target regions mediating ET recombination in the first round are shown in blue.

The region to be deleted in the second round is shown in orange. The selectable marker used in the second round was the gentamycin resistance gene (Gent), also shown in orange. The homology arms and target regions mediating ET recombination in the second round are shown in blue.

Mll exons 1, 11, 12 and 25 are represented by thin pink boxes. Intron 11, where the loxP-sA-IRES-βgeok-pA-loxP was targeted, is indicated by a lila bar.

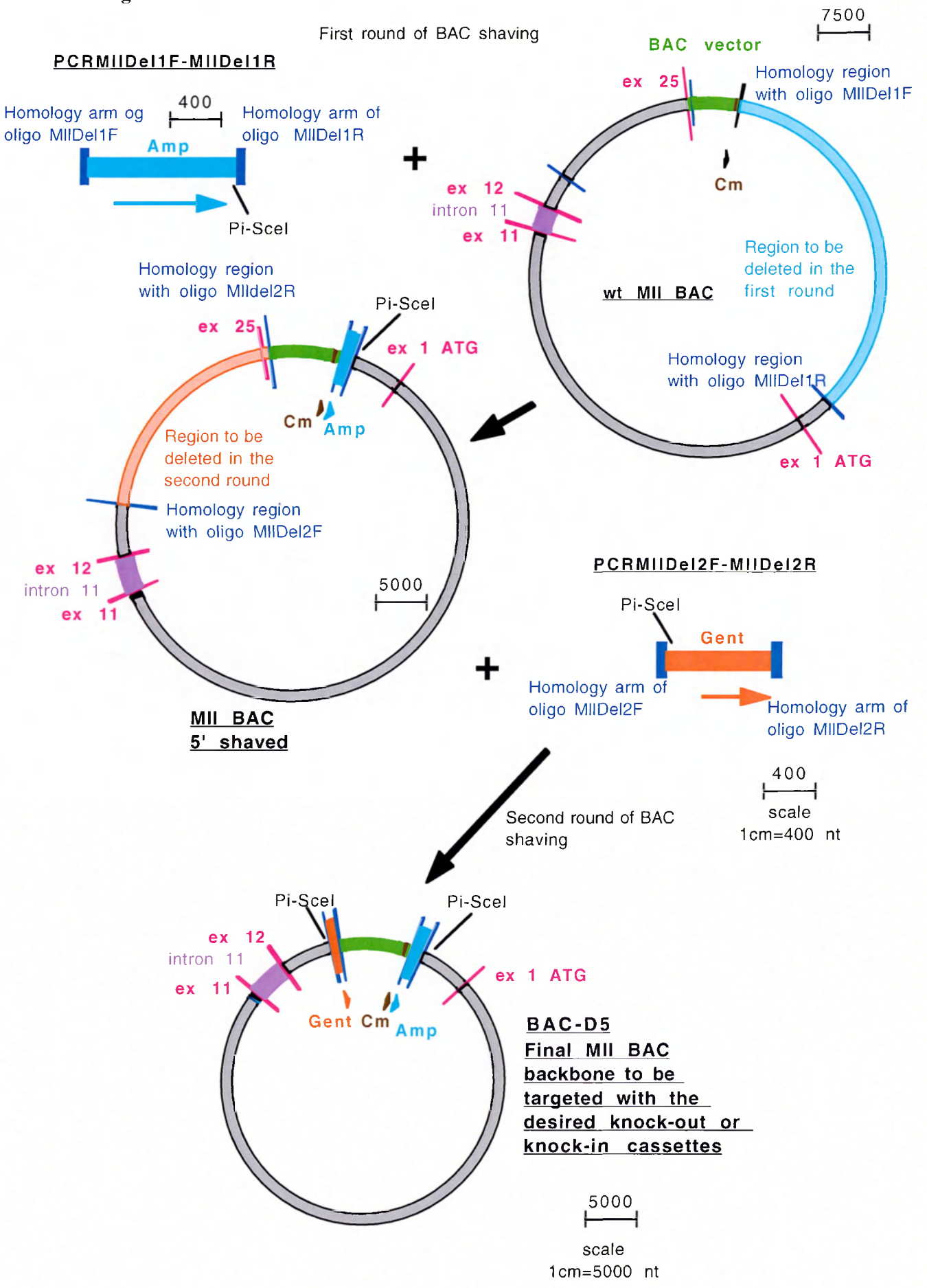
The scale bar indicates the number of nucleotides/centimeter.

Segments of the constructs accompanied by arrows indicate open reading frames.

Abbreviations: (Cm) chloramphenicol resistance gene; (Amp) βlactamase gene;

See text for a full description of the cloning strategy.

Figure 30


First round of BAC shaving
7500
BAC vector
ex 25
Homology region with oligo MIIDel1F
Cm
ex 12
intron 11
ex 11
Region to be deleted in the first round
wt MII BAC
Homology region with oligo MIIDel1R
ex 1 ATG
PCRMIIDel1F-MIIDel1R
400
Homology arm og oligo MIIDel1F
Amp
Homology arm of oligo MIIDel1R
Pi-SceI
Homology region with oligo MIIdel2R
Pi-SceI
ex 25
ex 1 ATG
Cm
Amp
Region to be deleted in the second round
Homology region with oligo MIIDel2F
ex 12
intron 11
ex 11
5000
MII BAC 5' shaved
PCRMIIDel2F-MIIDel2R
Pi-SceI
Gent
Homology arm of oligo MIIDel2F
Homology arm of oligo MIIDel2R
400
scale 1cm=400 nt
Second round of BAC shaving
Pi-SceI
Pi-SceI
ex 12
intron 11
ex 11
ex 1 ATG
Gent
Cm
Amp
BAC-D5
Final MII BAC backbone to be targeted with the desired knock-out or knock-in cassettes
5000
scale 1cm=5000 nt

enzyme PI-SceI (also called VDE). This site was included to cleave the final targeting construct from the BAC backbone. The portion of this oligo annealing to the β-lactamase gene (from plasmid pACYC177) comprises residues 103 to 132.

The PCR product (PCR*Mll*Del1F-*Mll*Del1R) was then electroporated into competent BAC cells from both batches n.3 and n.8, and cells were plated under double selection (chloramphenicol 12,5 μg/ml and ampicillin 50 μg/ml).

Fourteen colonies were analysed with BamHI. As shown in figure 31, all 14 colonies showed the intended deletion. However, colonies 1,3 and 10 also showed wild type bands, indicating that they were mixed colonies in which only a subset of cells had undergone ET deletion, conferring ampicillin resistance to the whole colony.

Two restriction patterns emerged from this first round of ET mediated deletion (colonies 1-11 versus colonies 12-14), differing by one band (a band of 9 kb is present in colonies 1 to 11 but absent from colonies 12 to 14, while a band of around 10 kb is present in colonies 12 to 14 but absent in colonies 1 to 11). Interestingly, colonies 1-11 had all originated from competent cells of batch n.3, while colonies 12-14 had originated from competent cells of batch n.8. This suggests that the intramolecular undesired recombination event had happened prior to the electroporation of the recombinogenic PCR product, either in the wild type strain HS996 as a result of intrinsic instability of this particular BAC, or during preparation of competent cells, when the induced expression of the recombinogenic proteins might have promoted intramolecular recombination between homologous regions of the BAC. To distinguish between these possibilities, and to exclude the possibility that the two patterns reflected two distinct deletion reactions mediated by the same PCR product (PCR*Mll*Del1F-*Mll*Del1R) due to multiple regions of homology, the following experiment was devised. A new PCR product was designed, whose homology arms to the *Mll* BAC lay just adjacent to the previous ones (employed for the PCR product PCR*Mll*Del1F-*Mll*Del1R).

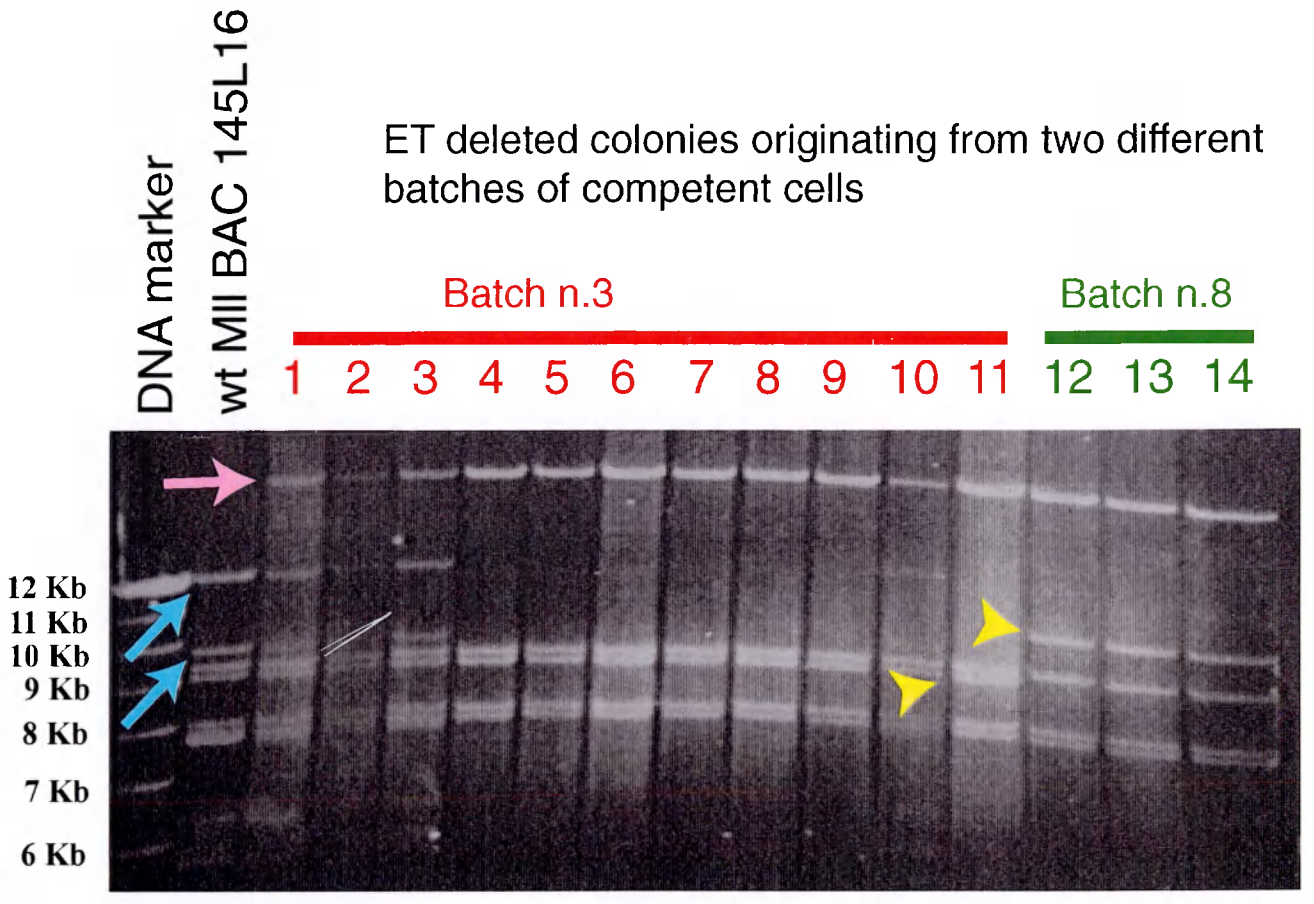


Figure 31 First round of MII BAC shaving (restriction analysis)
First round of ET mediated deletion ("BAC shaving") on the MII BAC 145L16 with the PCR product PCRMIIDEL1 conferring Ampicillin resistance. The light blue arrows indicate two of the bands which have been deleted from the wt BAC. The upper band (pink arrow) points to a new band which results from the deletion. Colonies 1, 2, 3 and 10 constitute mixed colonies which harbour both the wt and the deleted BAC. The pattern of colonies 4-11 and 12-14 differ for a single band (yellow arrowheads), which correlates with the original batch of competent cells utilised (see text and figures 31 and 32).

The assumption was that, if one of the patterns were resulting from inappropriate recombination with a region in the BAC other than the target one, after changing both homology arms, one should not observe this pattern again. In fact, the chances that both new homology arms would also have significant degrees of homology to the alternative spurious target regions were expected to be rather low. Plus, designing the new homology arms right next to the previous ones made it possible to directly compare the results of the two experiments, since the expected deletion pattern would not be affected by shifting the homology arms such a short distance. In order to completely exclude the role of sequences from the previous PCR product (PCR*Mll*Del1F-*Mll*Del1R) in the spurious deletion reaction, the selectable marker was also changed. The kanamycin resistance gene from the plasmid pACYC177 was chosen and amplified with the following oligonucleotides: oligo N*Mll*Del1F and oligo N*Mll*Del1R.

Oligo N*Mll*Del1F has the following5'-3' sequence:

CCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCG

GATAACAATTTAATTAATCTCTGATGTTACATTGCACAAGA.

Residues 1-64 constitute the arm of homology to the BAC backbone, in a position adjacent to the homology arm of oligo *Mll*Del1F. A PacI site was included at positions 65-72 to facilitate subsequent recognition of correctly targeted BACs. The portion of this oligo annealing to the kanamycin resistance gene (from plasmid pACYC177) comprises residues 73 to 96.

Oligo N*Mll*Del1R has the following 5'-3' sequence:

CACTGCAAATCAGTATCATGGTGTGACTAACATGGGTTCAGAGCCACCAGCCTA

GGCTTGATTTGCCCGGGCCCAGTGTTACAACCAATTAACCAATTC.

Residues 1-64 constitute the arm of homology to the *Mll* region 4968 bp upstream from the initiating ATG of the gene, adjacent to the homology arm of oligo *Mll*Del1R. Residues 65-

72 constitute a site for the rare cutting enzyme SrfI. The portion of this oligo annealing to the kanamycin resistance gene (from plasmid pACYC177) comprises residues 73 to 99. The resulting PCR product (PCRN*Mll*Del1F-N*Mll*DEl1R) was electroporated into the same batches of *Mll* BAC-containing competent cells (n.3 and n.8). BamHI digestion of the resulting colonies once again showed (figure 32) two distinct patterns, respectively associated with the batch of competent cells. In terms of efficiency, 5/6 colonies originating from batch n. 3 showed the correct deletion pattern (A,B,C,E and F). Colony n. D was an unrecombined colony which managed to survive selection. 3/6 colonies originating from batch n. 8 showed the correct deletion pattern (K, M and O). Colonies L, N and P were unrecombined colonies which also survived selection. The repeated occurrence of the same two patterns of deletion upon electroporation with two recombinogenic PCR products of completely unrelated sequence strongly suggest that the two patterns were preexisting, and were not the result of inappropriate recombination between the PCR product and spurious homologous regions in the BAC. Furthermore, comparison of colonies E, L, N and P (ie. unrecombined colonies respectively from batches n.3 and n.8) with the original wild type *Mll* 145L16 digest run alongside on the same gel as standard reference shows that batch n. 3 preserved the correct restriction pattern, while cells from batch n.8 did not. To confirm this, both batches n.3 and n.8 were streaked and individual colonies picked and digested with BamHI. In parallel, cells from the original glycerol stock of wt BAC 145L16 were also streaked, individual colonies picked and digested with BamHI. All three sets of digests were run alongside on a 0.4% agarose gel.

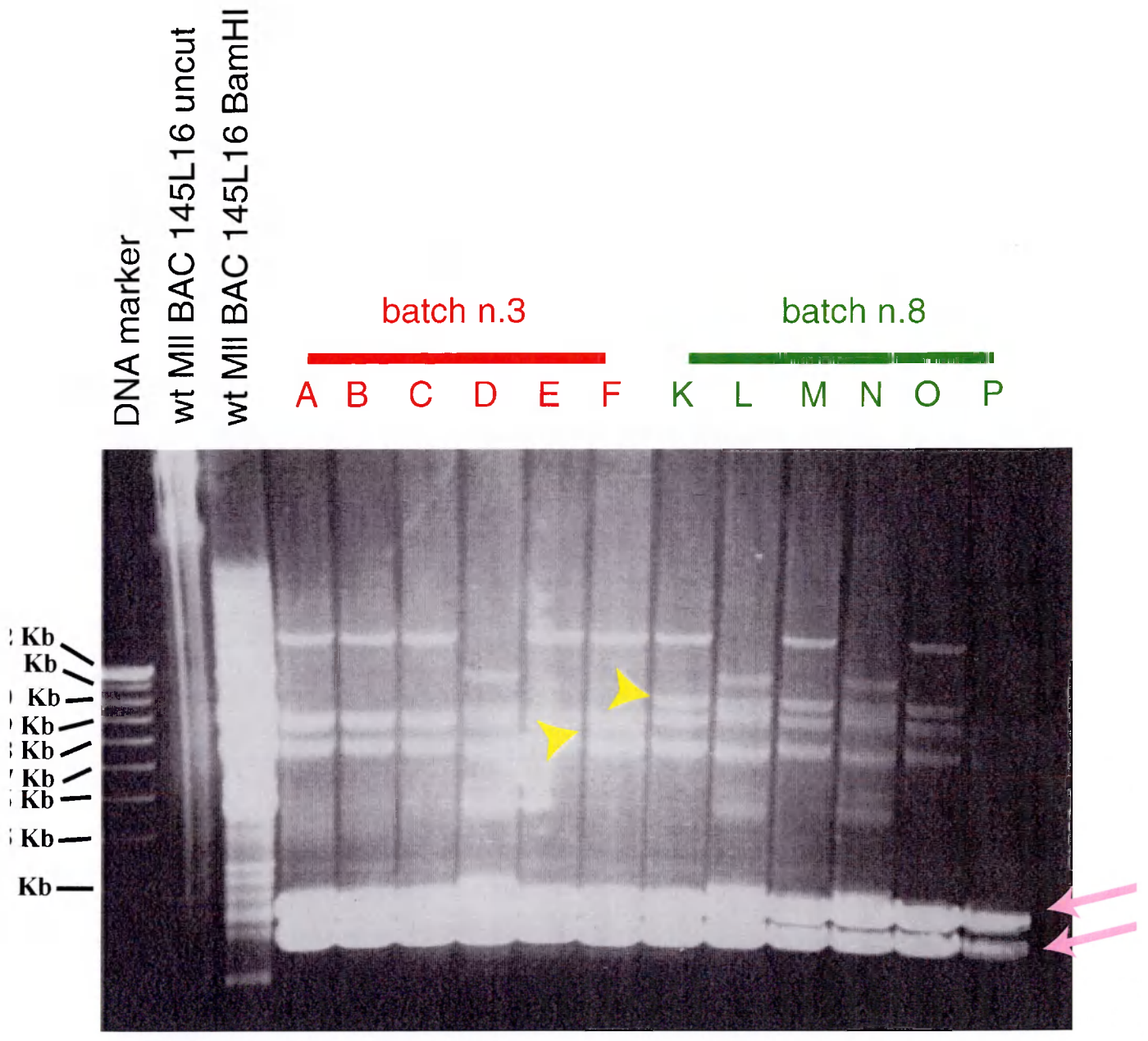


Figure 32 First round of Mll BAC "shaving" (control experiment)
ET mediated deletion of the Mll BAC with the PCR product NMllDel1 which confers kanamycin resistance and is unrelated to the PCR product MllDel1 used in the previous experiment (Figure 30) Results show that independent of the recombinogenic PCR product used, colonies originating from batch 3 and 8 of competent cells show consistently a single band difference in the restriction pattern (indicated by the yellow arrowheads). Colonies E, L, N and P represent unrecombined colonies which survived selection.
The two bands indicated by pink arrows originate from the recombinogenic plasmid R6Kαβγ which in this experiment was not lost from the cells after recombination.

The patterns (shown in figure 33) establish that all cells from batch n.3 are correct, while all cells from batch n.8 have undergone an intramolecular recombination event. The fact that all colonies from the original wt BAC show the same pattern (and none of them shows the pattern of batch n.8) supports the notion that the intramolecular recombination event probably occurred during preparation of competent cells upon expression of the recombinogenic proteins. This in turn probably resulted from the presence of significant homologous regions within the BAC.

In the second step of the deletion strategy, a PCR product was designed which would replace the 3' side of the *Mll* BAC from the BAC vector to a region approximately 5 kb downstream of the location in intron 11 that had been chosen for targeting the loxP flanked sA-IRES-βgeok-pA cassette. To this end PCR oligos *Mll*Del2F and *Mll*Del2R were synthethized to amplify the gentamycin resistance gene as a selectable marker replacing the 3' part of the BAC.

Oligo *Mll*Del2F has the following 5'-3' sequence:

CAGAGATAGGCGGATTTCTGAGTTCGAGGCCAGCCTGGTCTACAAAGTGAGATC

CAGGAGAGCCAAG*GCGCGCCATCTATGTCGGGTGCGGAGAAAGAGGTAATGAAATG*

*GCA***TGAAGGCACGAACCCAGTTGACATAAGCC.**

Residues 1-67 (underlined) constitute the homology arm to intron 16. Two rare restriction sites were included, an AscI site at positions 67-74 and a Pi-SceI site at positions 75-113 (italics). The portion of this oligo annealing to the gentamycin resistance gene comprises residues 114-142 (in bold).

Oligo *Mll*Del2R has the following 5'-3' sequence:

CTCTGTCGTTTCCTTTCTCTGTTTTTGTCCGTGGAATGAACAATGGAAGTCCGAG

CTCATCGCTA**TCGGCTTGAACGAATTGTTAGGTGGC**.

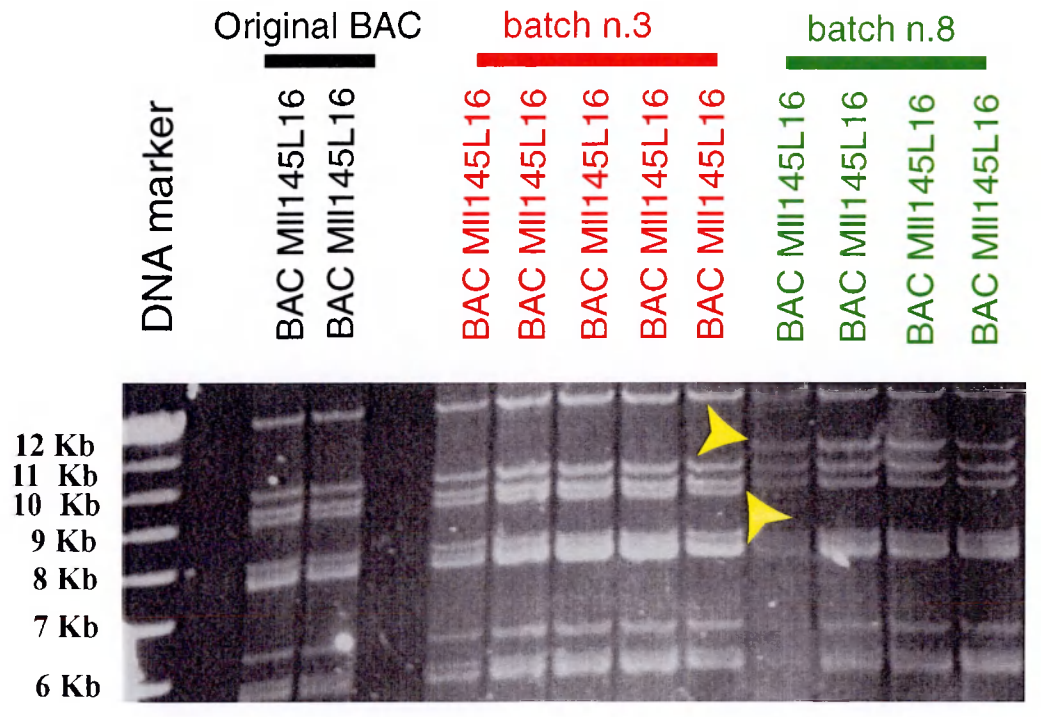


Figure 33 Restriction analysis of independent MII BAC colonies
BamHI digest of 2 independent colonies from a restreak of the original BAC containing cells, 5 independent colonies from a restreak of the batch of competent cells n.3, and 4 independent colonies from a restreak of the batch of competent cells n.8, in order to detect intramolecular recombination events prior to electroporation with the recombinogenic PCR product.
Cells coming from batch n. 3 and n. 8 display the same one band difference (yellow arrowheads) shown in previous experiments (Figure 31 and 32), demonstrating that the intramolecular recombination was preexistent to the transformation with the recombinogenic PCR product.

Residues 1-65 constitute the homology arm to the BAC backbone (underlined). The portion of this oligo annealing to the gentamycin resistance gene comprises residues 66-90 (in bold). Cells from colonies n.6 and n.7 of the first deletion step (figure 31) were grown in triple selection (chloramphenicol 12,5 μg/ml, ampicillin 50 μg/ml and tetracycline 25μg/ml) to maintain the BAC and the recombinogenic plasmid R6Kαβγ (which confers tetracycline resistance). After induction for 1 hr with L(+)arabinose, they were harvested at OD_{600} of 0.45 and made electrocompetent as described (see Materials and Methods). They were then electroporated with the PCR product PCR*Mll*Del2F-*Mll*Del2R and plated under triple selection (chloramphenicol 12,5 μg/ml, ampicillin 50 μg/ml and gentamycin 3μg/ml).

Twelve colonies were picked and analysed, 1-9 originating from colony n. 6, and 10-12 from colony n.7. From an electrophoresis analysis of their BamHI digests (Figure 34), two patterns clearly emerged, one represented by colonies 1, 5, 7 and 8, and the other represented by colonies 3, 9 and 12. Parallel analysis of the original colony n.6 (resulting from the first deletion step) showed that colony n. 7 was a mixed colony, in which some cells had not undergone this second round of deletion. Limited amounts of DNA for colonies 2, 6 and 11 prevented a clear assignment of their restriction pattern. Colony 10, on the other hand, showed a pattern distinct from either one of the two main ones, and since this pattern emerged only once, it was not further analysed. To distinguish which one of the two patterns reflected the correct deletion event, a sequencing analysis was performed. Sequencing primers were designed to display sequence around each end of the integration site, in intron 16 and in the BAC backbone. This established colony 5, and hence the most prevalent pattern also present in 1, 7 and 8, as correct. In colony 3, the PCR product had integrated at a site other than intron 16. Interestingly, sequences of colony 3 and 5 were identical at the 3' integration site, meaning that for colonies 3, 9 and 12, the PCR product had integrated

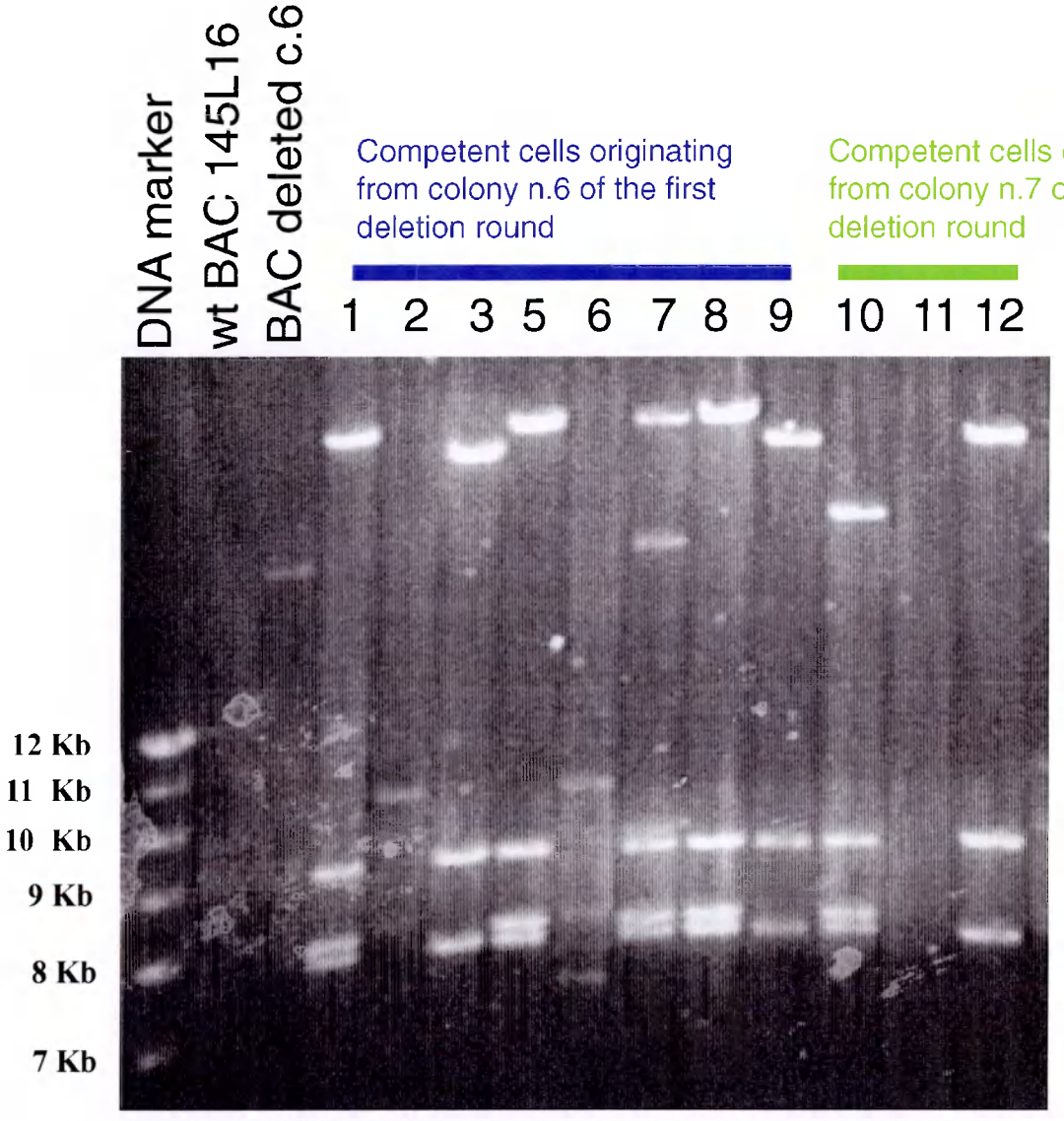


Figure 34 Second round of Mll BAC shaving (restriction analysis)
BamHI digest on candidate recombined colonies. Competent cells were prepared from colony n.6 and n.7 resulting from the first deletion round and were then electroporated with the PCR product MllDel2. Two patterns emerged, one represented by colonies 1, 5, 7 and 8, and the other represented by colonies 3, 9 and 12.
Sequencing analysis established the pattern displayed by colonies 1, 5, 7 and 8 as the correct one.

correctly with the 3' homology arm, but incorrectly with the 5' one at presumably the same place. Thus, contrary to the case discussed above, in this instance, one of the two restriction patterns could actually be explained by mistargeting of the PCR product,possibly due to the presence of spurious sites with relevant homology to the sequence of the homology arms.

On the basis of these results, the deleted BAC from colony n. 5 (BACD-5) now constituted the final backbone into which the two functional knock-in cassettes had to be targeted, the TAP-FRT-pA-hygro-PGK-FRT cassette immediately downstream of the initiating ATG, and the loxP-sA-IRES-βGeok-pA-loxP cassette in intron 11.

IX.3.4 Assembly of the knock-in cassette *Mll*-TAP-hygro

An outline of the cloning steps for the generation of the *Mll*-TAP-hygro cassette is shown in figure 35. The knock-in cassette *Mll*-TAP-hygro consists of the following elements:

1) The 5' *Mll* homology arm (of 49 bp) including the starting ATG of the *Mll* gene and the 46 nucleotides upstream of it.
2) The TAP (Tandem affinity purification) protein tag (see above), which is a fusion of the Protein A, the TEV cleavage site and the Calmodulin Binding Domain (CBD)
3) A 5' FRT site, the target site for the Flp site specific recombinase
4) A module expressing the hygromycin phosphotransferase gene (which confers resistance to hygromycin) under the control of both the PGK promoter (for selection in mammalian cells) and a prokaryotic promoter (for selection in *E. coli*). This module is placed in an orientation opposite to the direction of transcription of the *Mll* gene, in order to minimise the possibility of promoter interference between the endogenous *Mll* promoter and the PGK promoter.
5) A 3' FRT, which, together with the 5' FRT, allows excision of the hygromycin expressing module upon expression of Flp recombinase. The 3' FRT is followed by a single nucleotide inserted before the 3' *Mll* homology arm, so that, after excision of the hygromycin module, the reading frame is maintained between the starting ATG, the TAP tag, the FRT (left after the recombination reaction) and the rest of *Mll* exon 1, including the 3' *Mll* homology arm.
6) The 3' *Mll* homology arm, which includes the second codon of *Mll* exon 1 and 66 nucleotides downstream of it. Together with the 5' *Mll* homology arm, it directs the insertion of the cassette between the first and the second codon of *Mll* exon 1.

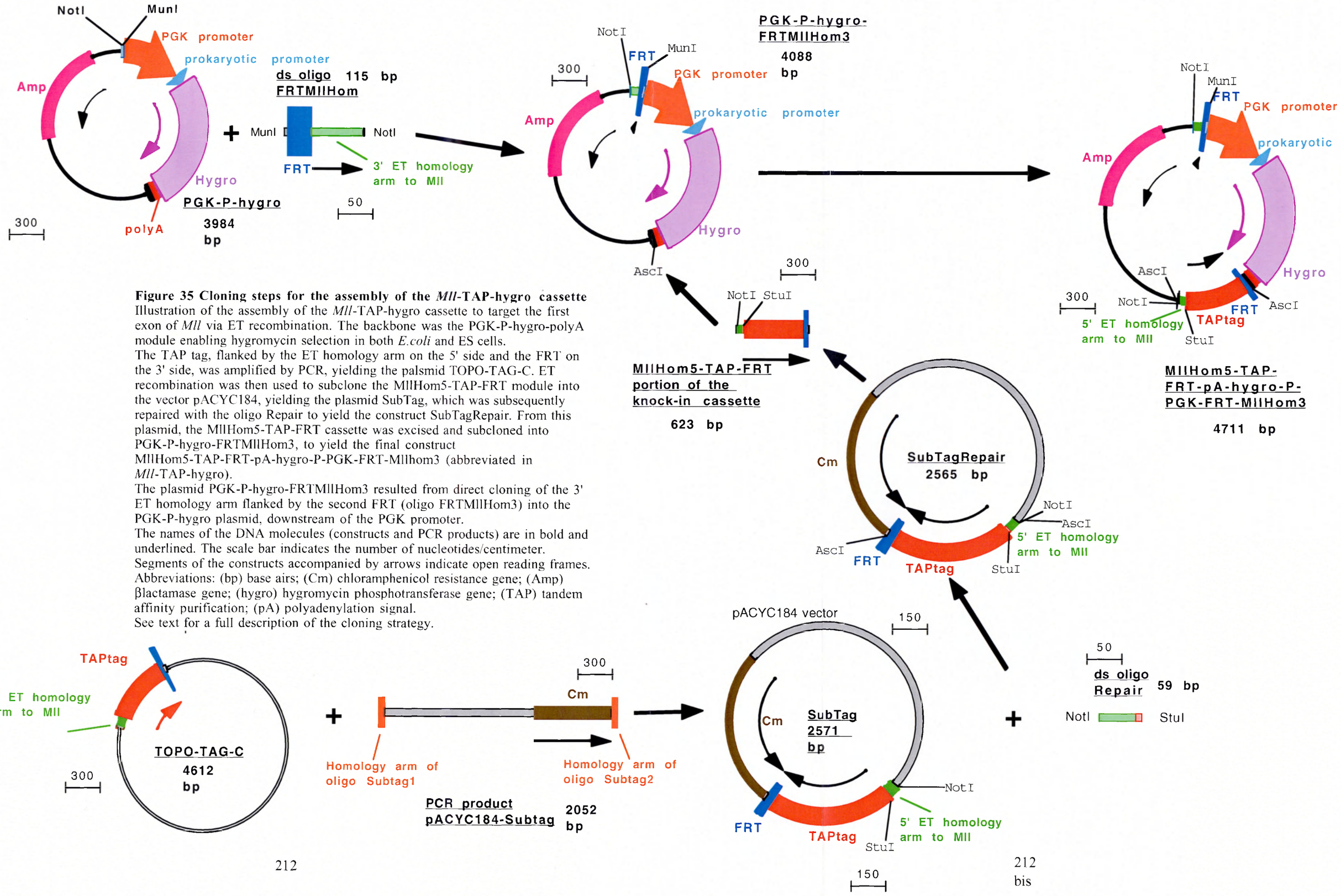


Figure 35 Cloning steps for the assembly of the *Mll*-TAP-hygro cassette
Illustration of the assembly of the *Mll*-TAP-hygro cassette to target the first exon of *Mll* via ET recombination. The backbone was the PGK-P-hygro-polyA module enabling hygromycin selection in both *E.coli* and ES cells.
The TAP tag, flanked by the ET homology arm on the 5' side and the FRT on the 3' side, was amplified by PCR, yielding the palsmid TOPO-TAG-C. ET recombination was then used to subclone the MllHom5-TAP-FRT module into the vector pACYC184, yielding the plasmid SubTag, which was subsequently repaired with the oligo Repair to yield the construct SubTagRepair. From this plasmid, the MllHom5-TAP-FRT cassette was excised and subcloned into PGK-P-hygro-FRTMllHom3, to yield the final construct MllHom5-TAP-FRT-pA-hygro-P-PGK-FRT-Mllhom3 (abbreviated in *Mll*-TAP-hygro).
The plasmid PGK-P-hygro-FRTMllHom3 resulted from direct cloning of the 3' ET homology arm flanked by the second FRT (oligo FRTMllHom3) into the PGK-P-hygro plasmid, downstream of the PGK promoter.
The names of the DNA molecules (constructs and PCR products) are in bold and underlined. The scale bar indicates the number of nucleotides/centimeter.
Segments of the constructs accompanied by arrows indicate open reading frames.
Abbreviations: (bp) base airs; (Cm) chloramphenicol resistance gene; (Amp) βlactamase gene; (hygro) hygromycin phosphotransferase gene; (TAP) tandem affinity purification; (pA) polyadenylation signal.
See text for a full description of the cloning strategy.

The starting backbone for the assembly of this cassette was the plasmid PGK-P-hygro (figure 35), in which a prokaryotic promoter (P) had been placed immediately downstream of the PGK promoter, to permit hygromycin selection both in eukaryotic and prokaryotic systems. This is particularly useful, since it enables one to use the same selectable marker, and hence the same cassette, for ET mediated homologous recombination in *E. coli* and for homologous recombination in mouse ES cells.

The first step involved cloning, immediately downstream of the PGK promoter (since the hygromycin module had to be oriented opposite to the direction of transcription of the *Mll* gene) the 3' *Mll* homology arm, flanked, on its 5' side, by an FRT element. To this end, the following oligos were synthethized: FRT*Mll*HomF and FRT*Mll*HomR, which would anneal to each other so as to create an MfeI site on the 5' end and a NotI site on the 3' end for cloning in the corresponding sites of PGK-P-hygro.

Oligo FRT*Mll*HomF has the following 5'-3' sequence:

AATTG**GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTC***A*GCGCACAGCTGT CGGTGGCGCTTCCCCGCCCGACCCGGGACCACCGGGGGCGGCGGCGGCGGGGG GCGCCGGGGC.

The FRT is located at positions 6 to 39 (in bold).

Residues 41 to 114 (underlined) constitute the region of homology to codons 2 through 26 of *Mll* exon 1. A single nucleotide was inserted at position 40 (italics) in order to keep the reading frame between the starting ATG, the TAP tag, the FRT (left after the recombination reaction) and the rest of *Mll* exon 1 after excision of the hygromycin module. The sequence AATTG (positions 1-5) provided the upper strand of the MfeI site, while the the terminally located GC dinucleotide (positions 114-115) formed the upper strand of the NotI site.

Oligo FRT*Mll*HomR has the following 5'-3' sequence:

GGCCGCCCCGGCGCCCCCCGCCGCCGCCGCCCCCGGTGGTCCCGGGTCGGGCG
GGGAAGCGCCACCGACAGCTGTGCGCTGAAGTTCCTATACTTTCTAGAGAATAG
GAACTTCC.

It is complementary to FRT*Mll*HomF throughout residues 5 to 114. Upon annealing with FRT*Mll*HomF, the GGCC sequence (positions 1-4) constitutes the lower strand of the NotI site, while the terminally located C (position 115) forms the lower strand of the MfeI site.

The two oligos were annealed to each other and then ligated to the vector PGK-P-hygro, which had been previously digested with MfeI and NotI. Cells transformed with the ligation mixture were grown on double selection (ampicillin 100μg/ml and hygromycin 50 μg/ml) All colonies analysed showed the correct pattern of integration of the FRT*Mll*Hom insert into the PGK-P-hygro vector. It has been reported, and it was also observed in some of these cloning procedures, that long oligonucleotides (like the ones used in this case) can sometimes contain defects (mostly depurination) which can result in deletion of the affected residues upon scrutiny by the E.coli DNA repair machinery. Since in this case it was absolutely crucial that the sequence of FRT*Mll*Hom be correct, in order to keep the reading frame and to maintain the functionality of the FRT site, some recombinant colonies (PGK-P-hygro-FRT*Mll*Hom3) were sequenced. Colony n. 2 was found to be correct, and was therefore chosen for the subsequent cloning steps.

The problem of mutations and defects in long oligos (particularly with a high GC content) was particularly troublesome for the assembly of the 5*Mll*Hom-TAP-FRT portion of the *Mll*Hom5-TAP-FRT-pA-hygro-P-PGK-FRT-*Mll*Hom3 cassette. Once again, an ET cloning based subcloning approach (see below) eventually led to the correct clone.

As a starting point, the TAP tag was amplified from a template by PCR with the following oligos: New*Mll*NTagF and NtagR.

Primer New*Mll*NtagF has the following 5'-3' sequence:

TTTGGCGCGCCATTGCGGCCGCTCCCCCCCCTCCGCCTCCCCGCCCCCCTGTGTT
GTCGCCTCTCCCTCTCGCTGCTTCACTTACGGGGCGAACATGCGCGCCGCAGGC
CTTGCGCAACACGATGAAGCCGTGGACAACAAATTCAAC.

Residues 18-94 constitute the arm of homology to the region of *Mll* immediately upstream of the initiating ATG, which is also included (positions 95-97). The portion of the oligo which anneals to the TAP containing template comprises residues 98 through 148. An AscI site (positions 4-11) and a NotI site (positions 15-22) were inserted for further cloning steps and, for the NotI site, to release the recombinogenic cassette from the vector backbone prior to ET recombination.

Oligo NewNTagR has the following 5'-3' sequence:

TACAGGCGCGCCGGAAGTTCCTATACTTTCTAGAGAATAGGAACTTCCATCAAG
TGCCCCGGAGGATGAGATTTTC.

Residues 14 through 47 constitute the FRT site, while the portion of the oligo which anneals to the TAP containing template comprises residues 49 through 76. A single residue (a C at position 48) was included in order to maintain the reading frame between the TAP tag anf the FRT. An AscI site is present at residues 5-12.

The PCR product was cloned into a PCR-TOPOII vector (commercially available from Invitrogen) and transformed into TOP10 competent cells. In this experiment, two elements were a potential source of mutation: defects in oligonucleotide synthesis and the mutation rate intrinsic to any PCR reaction. Although a high fidelity Taq polymerase was used, this risk can never be fully eliminated. Therefore, of 20 colonies shown to contain an insert of the correct size, 8 colonies were sequenced to check the integrity of the 5' Mll homology arm, the TAP tag and the FRT. None of the colonies had the absolutely correct sequence, and most mutations were detected in the oligonucleotide portions of the PCR product, arguing that their most likely source was a defect during oligonucleotide synthesis. One

colony (TOPOTAG-C) was almost entirely correct, except for a G to A mismatch at position 49 (corresponding to position 59 on the oligo New*Mll*NtagF), and the missed incorporation of residues 1-10 of the oligo NewNTagR (including the AscI site) in the cloned PCR product. Given the seemingly high rate of mutations occurring in these oligonucleotides (regardless of any purification technique employed), and since this colony contained almost the full correct sequence, I decided to use an ET subcloning approach to repair these mutations without having to undergo a further round of PCR amplification. The strategy involved subcloning the correct relevant part of TOPOTAG-C into an acceptor plasmid (pACYC184) while at the same time incorporating in the ET oligonucleotides the parts which were missing or mutated in TOPOTAG-C. OligosSubtag1 and Subtag2 were designed:

Oligo Subtag1 has the following 5'-3' sequence:

GCGCATGTTCGCCCCGTAAGTGAAGCAGCGAGAGGGAGAGGCGACAACACAGG CGGCCGCAAAGGCGCGCCACAACTTATATCGTATGG.

Residues 1-53 constitutes the region of homology to the corresponding region of TOPOTAG-C. The two sequences are completely identical, except that in the oligo Subtag1, a C is inserted at position 42 in order to correct the G to A mismatch in TOPOTAG-C. Furthermore, in the new configuration resulting from this subcloning step, the actual 5' *Mll* homology arm has been reduced to 48 nucleotides. This has been shown to be sufficient to promote homologous recombination in E.coli; in this case, this shortening excluded from the Subtag1 oligo the GC rich stretch (GCCGCTCCCCCCCCTCCGCCTCCCCGCCCC), which was present in the oligo New*Mll*NtagF and which could have created further problems during the oligo synthesis.

The portion of this oligo which anneals to the pACYC184 template to amplify the chloramphenicol resistance gene comprises residues 72 to 89. An AscI site was inserted

(positions 64-71) for the next cloning step, as well as a NotI site (positions 53-60) to release the recombinogenic *Mll*-TAP-hygro cassette from the vector backbone prior to ET recombination.

Oligo SubtagR has the following 5'-3' sequence:

GATGGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCCGGCGCGCCTTACGCC CCGCCCTGCCACTC.

Residues 1-41 constitute the arm of homology to the corresponding region in TOPOTAG-C. Residues 42-47 (CGCGCC) were inserted in order to reconstitute the AscI site, which had been deleted in TOPOTAG-C. The portion of this oligo which anneals to the pACYC184 template to amplify the chloramphenicol resistance gene comprises residues 48 to 68.

The PCR product features at its ends the two regions of homology to the 5'*Mll*Hom-TAP-FRT module of TOPOTAG-C (acting as the donor plasmid).

The PCR product was first digested with DpnI, an enzyme which cleaves only when its restriction site (GATC) is methylated, and thereby serves after the PCR to eliminate the template (in this case pACYC184), which could otherwise contribute a serious source of antibiotic resistant colonies. Following purification of the DpnI digest, the PCR product (300 nanograms) was coelectroporated together with the donor plasmid (TOPOTAG-C) into the E.coli strain JC8679, which constitutively expresses the recombinogenic proteins RecE and RecT. Cells were then plated on chloramphenicol selection (50 μg/ml).

Eight out of ten colonies analysed displayed the correct pattern. Colony n.1 (Subtag1) was selected for further cloning.

In colony SubTag1, six residues (CGCGCC) occurred immediately downstream of the initiating *Mll* ATG, coding for the two amino acids arginine and alanine. This stretch of sequence had been incorporated in the oligo New*Mll*NtagF in order to amplify a previous version of the TAP tag containing these two additional amino acids. However, in the final

version of the TAP tag, these two amino acids are absent, and it was therefore important to remove them from the plasmid Subtag1, since their effect on the overall structure of the tag could not be foreseen. To this end, an oligo repair strategy, which relied on the presence of unique NotI and StuI sites flanking the region to be modified, was adopted. The RepairF and RepairR oligos were synthethized, complementary to each other. Upon annealing, they create on the 5' side a 5' overhang compatible with a NotI site, and at the 3' a blunt end compatible with an StuI site.

Oligo repair F has the following 5'-3' sequence:

GGCCGCCTGTGTTGTCGCCTCTCCCTCTCGCTGCTTCACTTACGGGGCGAACATG GCAGG.

The sequence is identical to the corresponding stretch present in Subtag1, except that residues 53 through 60 (ATGGCAGG) replace the longer portion (ATGCGCGCCGCAGG) present in SubTag1 harbouring the two incorrect residues arginine and alanine.

Oligo RepairR has the following 5'-3' sequence:

CCTGCCATGTTCGCCCCGTAAGTGAAGCAGCGAGAGGGAGAGGCGACAACACA GGC.

The two oligos were annealed to each other and then ligated to the vector SubTag1, which had been previously digested with StuI and NotI. Cells transformed with the ligation mixture were grown on double selection (chloramphenicol 50 μg/ml) All colonies analysed (SubTagRepair 1-6) showed the correct pattern of integration of the Repair oligo insert into the SubTag1 vector.

Now, the SubtagRepair plasmid harboured the correct version of the 5*Mll*Hom-TAP-FRT portion of the knock-in cassette *Mll*Hom5-TAP-FRT-pA-hygro-P-PGK-FRT-*Mll*Hom3. To finally assemble the whole knock-in cassette *Mll*Hom5-TAP-FRT-pA-hygro-P-PGK-FRT-*Mll*Hom3, the 5*Mll*Hom-TAP-FRT was excised from the SubtagRepair plasmid by AscI

digestion, and subcloned by conventional ligation into PGK-P-hygro-FRT*Mll*Hom3 colony n.2 (see above). 4 out of 8 colonies had integrated the insert in the correct orientation. Colonies n. 5 and n. 8 were sequenced and found to be correct. A maxiprep DNA preparation of colony n. 8 was digested with NotI and AseI to perform the ET cloning recombination step on the *Mll* shaved backbone. The NotI digest released the recombinogenic insert *Mll*Hom5-TAP-FRT-pA-hygro-P-PGK-FRT-*Mll*Hom3 from the vector backbone, while the AseI site had the purpose of further cleaving the vector backbone which could have constituted a source of background upon religation.

IX.3.5 Assembly of the knock-in cassette loxP-βGeok-loxP

As a preliminary step to the generation of the knock-in cassette loxP-sA-IRES-βGeok-pA-loxP, a combination of PCR-mediated cloning and sequencing of relevant introns as well as direct BAC sequencing were utilised to sequence the mouse *Mll* region spanning exon 10 to 19. This served the purpose of identifying the target region in which to place the loxP flanked knock-in cassette, as well as to derive sufficient sequence information for the development of a Southern strategy to screen recombinant ES cells.

The mouse sequence from exon 10 to 19 turned out to be 10778 bp long. Intron 11 was chosen for the following reasons. First, it was known to be one of the most common *Mll* introns disrupted by translocations. Furthermore, previous mouse models, aimed at recapitulating the *Mll-Af9* translocation, had already targeted the gene at this site. The genomic breakpoints of three patients carrying *Mll* translocations have been completely sequenced, and two of them interrupt intron 11. However, comparison between human and mouse sequence in intron 11 did not identify any region of high similarity, providing no specific indication as to where the loxP site would best be placed within intron 11.

The backbone for the assembly of the knock-in cassette loxP-sA-IRES-βGeok-pA-loxP was the previously described vector PUX4-sA-IRES-βGeok-pA. As in the *Af4* case outlined above, direct PCR amplification of the sA-IRES-βGeok-pA cassette with oligos harbouring regions of homology to the target sites in intron 11, while theoretically possible, would have meant an unacceptably high risk of mutations in functionally relevant regions of the cassette, which would not be detected prior to the ES cells or even the mouse experiment. Therefore, a strategy was devised, analogous to the one already described for the generation of the *Af4* allele. Homology arms were generated by PCR, followed by restriction digest and conventional ligation into the two ends of the targeting vector, using unique restriction sites. This assured that the core of the targeting construct was never subjected to PCR amplification.

The 5' homology arm was amplified by PCR from wild type *Mll* BAC DNA using the oligos N*Mll*Hom1F and *Mll*Hom1R. Forward oligo N*Mll*Hom1F has the following 5'-3' sequence:

TTATACCATGGCGCGCCGGCCGGCCACAGCTTTCATCTCCGCATTC.

At position 10-17, an AscI site was incorporated for further cloning, while the annealing portion includes residues 26-46. Reverse oligo *Mll*Hom1R has the following 5'-3' sequence:

CTGAGGTGCCTTAATTAAATAACTTCGTATAGCATACATTATACGAAGTTATTTT

TATTTATGTGTATTATCCTC.

At position 11-18, a PacI site was incorporated for further cloning; the loxP site is located at position 19-52, while the annealing portion includes residues 53-75. The resulting PCR product (PCR*Mll*Hom1), which is 355 nucleotides long and contains 277 nucleotides of sequence identity with the region of the BAC to be targeted, was cloned into PUX4-sA-IRES-βGeok-pA by AscI and PacI ligation. All colonies showed the correct pattern and colony 6 was chosen for the subsequent cloning step (PUX4-*Mll*Hom1-sA-IRES-βGeok-pA c6.)

The 3' homology arm was amplified from *Mll* BAC DNA using the oligos N*Mll*Hom2F and *Mll*Hom2R. Forward oligo N*Mll*Hom2F has the following 5'-3' sequence:

GGCAATCTTCGGTTTAAACATAACTTCGTATAATGTATGCTATACGAAGTTATGG

CGCGCCGCCCGGGCTAATTATGACCACTGTTTGAGGA.

A PmeI site (position 12-19) was incorporated for further cloning. The loxP site is located at position 20-53, while the annealing portion includes residues 70-92. A rare cutter polylinker, containing restriction sites for AscI (54-61) and SrfI (63-69) was inserted for the pulsed-field-gel-electrophoresis Southern strategy, in order to check whether the whole *Mll* targeting construct had integrated on the same allele (see above). Reverse oligo *Mll*Hom2R has the following 5'-3' sequence:

ACTAAGATTCCGCGGTGGCGGCCGCTGACGATACAGAGGTTACACAAG.

SacII (10-15) and NotI (17-25) were inserted for further cloning. The annealing portion includes residues 26-48. The resulting PCR product (PCR*Mll*Hom2), which is 412 base pairs long and contains 333 residues of sequence identity with the region of the BAC to be targeted, was cloned into PUX4-*Mll*Hom1-sA-IRES-βGeok-pA c6. All colonies showed the correct pattern. To check the integrity of the loxP sites, colony 2 (PUX4-*Mll*Hom1-sA-IRES-βGeok-pA-*Mll*Hom2c.2) was sequenced, found to be correct, and was selected for the final ET cloning step into the *Mll* shaved BAC backbone.

IX.3.6 Targeting of the knock-in cassette loxP-βGeok-loxP into the *Mll* BAC shaved backbone

To generate the final ES targeting construct TAP-*Mll*-LacZ, the shaved BAC backbone BACD-5 was sequentially targeted with the loxP-sA-IRES-βGeok-pA-loxP and the *Mll*Hom5-TAP-FRT-pA-hygro-P-PGK-FRT-*Mll*Hom3 cassettes (figure 36). In order to target the knock-in cassette loxP-sA-IRES-βGeok-pA-loxP, BAC-D5 containing cells were made electrocompetent (see Materials and Methods). They were then electroporated with about 500 ng of the knock-in cassette loxP-sA-IRES-βGeok-pA-loxP and plated on quadruple selection (cholramphenicol 12,5μg/ml, ampicillin 50 μg/ml, gentamycin 3 μg/ml and kanamycin 20 μg/ml).

In order to screen for the correct recombinant colonies a PCR strategy was devised in which the two primers would anneal across the predicted integration site. The following primers were synthethized: *Mll*PUXF and *Mll*i11R.

Primer *Mll*PUXF has the following 5'-3' sequence:

AAACCTCCCACACCTCCCCCTGAA.

Primer *Mll*i11R has the following5'-3' sequence:

CACAGAGAAATAATAACCATCGTC

Nine out of ten colonies, screened by colony PCR, showed a band of the correct size (Figure 37). While such a PCR strategy is a fast and efficient way to identify colonies which have probably undergone the correct recombination reaction, the restriction patterns of the candidate colonies must still be analysed by agarose gel electrophoresis, in order to exclude the presence of undesired intramolecular recombination events. As outlined above, this is rather straightforward. First, if one analyses a sufficient number of independent recombinant colonies, in most cases a unique, or at least a predominant pattern emerges which enables one to exclude colonies showing a clearly aberrant pattern. In those cases in which two

Figure 36 Sequential targeting of the shaved Mll BAC

The shaved Mll BAC (BAC-D5) was targeted sequentially with the loxP-flanked βGeok cassette (loxP-sA-IRES-βgeok-loxP) in intron 11 and with the FRT-flanked TAP-Hygro cassette (MllHom5-TAP-FRT-pA-hygro-P-PGK-FRT-Mllhom3) downstream of the intiating ATG, to yield the final targeting construct TAP-Mll-LacZ.

The 5' and 3' homology arms mediating ET recombination are shown in green for both cassettes.

The names of the DNA molecules (constructs and PCR products) are in bold and underlined. The scale bar indicates the number of nucleotides/centimeter.

Segments of the constructs accompanied by arrows indicate open reading frames.

Abbreviations: (Cm) chloramphenicol resistance gene; (Amp) βlactamase gene; (hygro) hygromycin phosphotransferase gene; (TAP) tandem affinity purification; (Gent) gentamycin resistance gene; (ECMV IRES) encephalomyocarditis virus internal rybosomal entry site; (LacZ-neo) fusion of the β-galactosidase and the neomycin genes; (pA) polyadenylation signal; (SV40) simian virus 40;

See text for a full description of the cloning strategy.

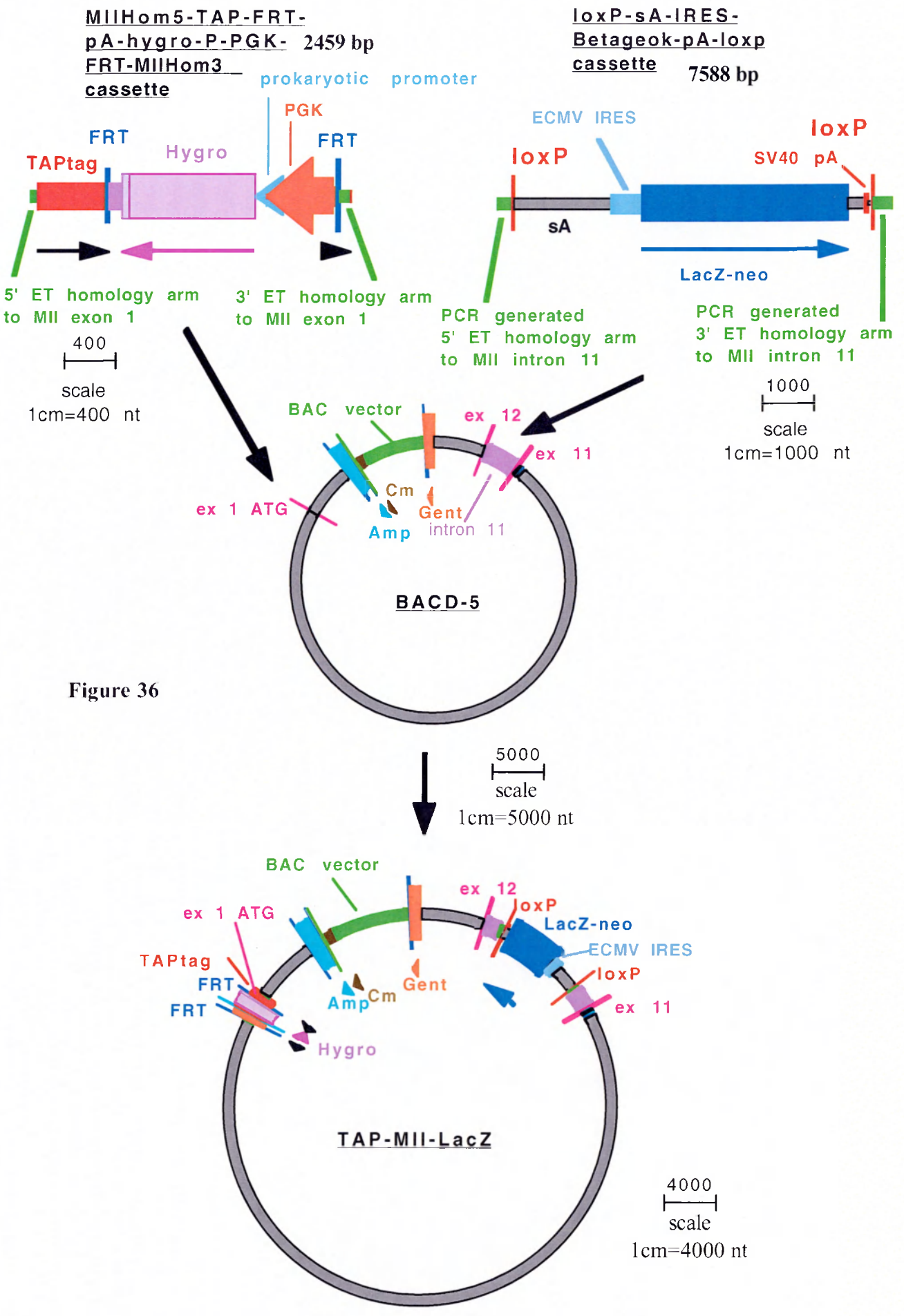

MIIHom5-TAP-FRT-pA-hygro-P-PGK-FRT-MIIHom3 cassette
2459 bp
prokaryotic promoter
PGK
FRT
Hygro
TAPtag
FRT
5' ET homology arm to MII exon 1
3' ET homology arm to MII exon 1
400
scale
1cm=400 nt
loxP-sA-IRES-Betageok-pA-loxp cassette
7588 bp
ECMV IRES
loxP
SV40 pA
loxP
sA
LacZ-neo
PCR generated 5' ET homology arm to MII intron 11
PCR generated 3' ET homology arm to MII intron 11
1000
scale
1cm=1000 nt
BAC vector
ex 12
ex 11
Cm
Gent
ex 1 ATG
Amp
intron 11
BACD-5
5000
scale
1cm=5000 nt
BAC vector
ex 12
loxP
ex 1 ATG
LacZ-neo
ECMV IRES
TAPtag
loxP
FRT
Gent
Amp
Cm
ex 11
FRT
Hygro
TAP-MII-LacZ
4000
scale
1cm=4000 nt

Figure 36

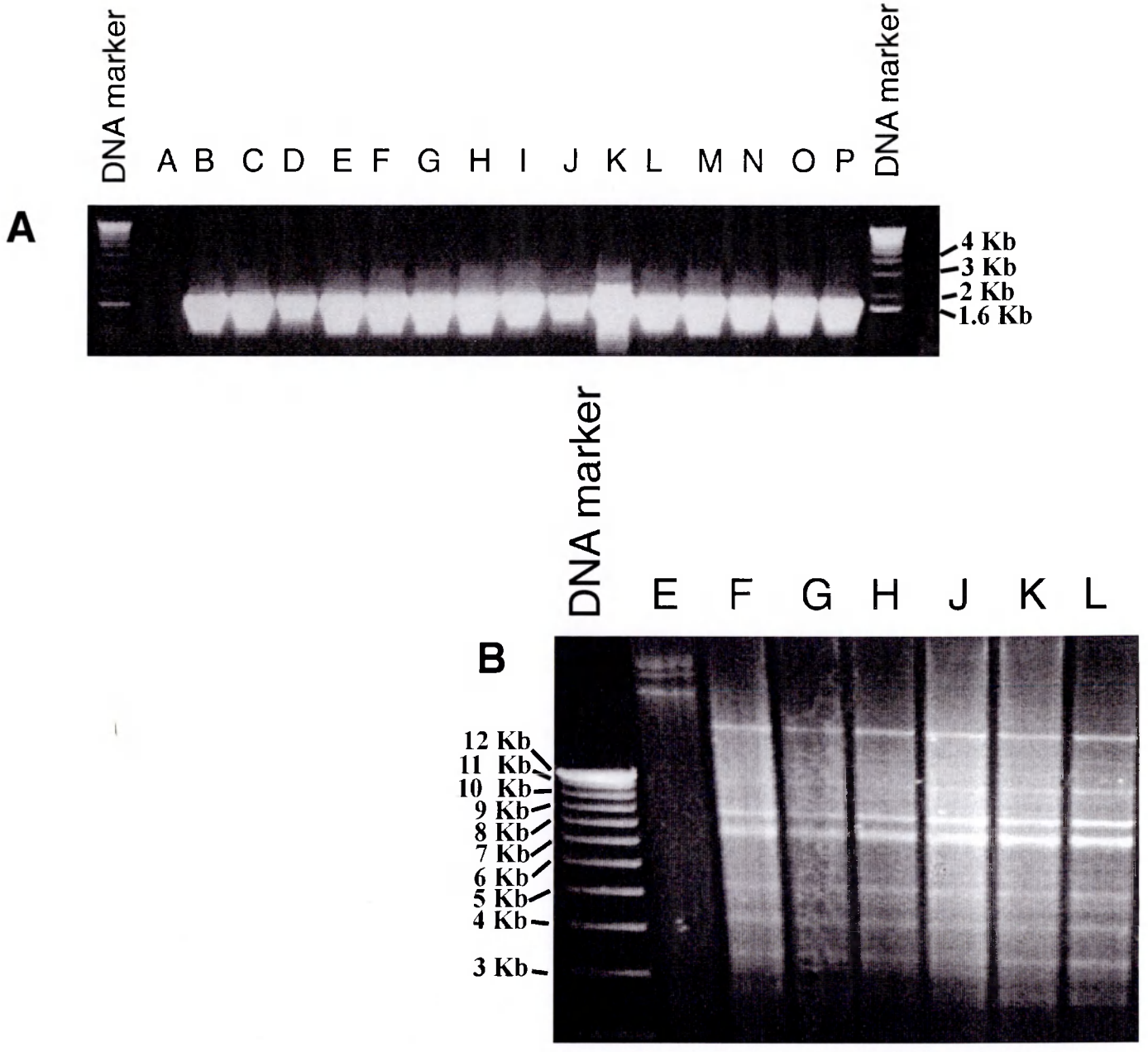


Figure 37 Targeting of the LoxP-flanked βGeok cassette to the Mll shaved BAC backbone
The loxP-flanked βGeok cassette was targeted by ET recombination to the intron 11 of the Mll gene on the "shaved" BAC backbone resulting from the two rounds of ET mediated deletion (Figures 29-33).
A. PCR screening of candidate colonies with primers spanning the integration site.
B. Representative BamHI digests of PCR-positive colonies. All colonies examined except colony E showed the same restriction pattern

equally frequent patterns emerge, as in the cases of the first and second deletion rounds described above, the correct colonies will usually become apparent by comparison with the original wild type BAC and the BACs derived from the previous ET modifications. To this end, BamHI digests were carried out on the 9 colonies which had been positive by PCR. All of them showed the same restriction pattern, indicating that the ET targeting had occurred in the same position in all cases. Sequencing was performed on colonies B and J, and confirmed correct integration site in intron 11. Therefore, these two colonies (*Mll*BACD5-loxP-sA-IRES-βGeok-pA-loxP n. B and J) were selected for the final step in the assembly of the *Mll* targeting construct (ie. targeting of the *Mll*Hom5-TAP-FRT-pA-hygro-P-PGK-FRT-*Mll*Hom3 knock-in cassette).

IX.3.7 Targeting of the knock-in cassette *Mll*-TAP-hygro into *Mll* BAC shaved backbone

Cells harbouring the *Mll*BACD5-loxP-sA-IRES-βGeok-pA-loxP BAC (from both batches B and J) were transformed with the recombinogenic plasmid R6K-αβγ plasmid, plated on quintuple selection (chlramphenicol 12,5μg/ml, ampicillin 50 μg/ml, gentamycin 3 μg/ml, kanamycin 20 μg/ml and tetracycline 25 μg/ml), and the correct transformants identified by agarose gel electrophoresis to detect the presence of the recombinogenic plasmid. Cells were then made electrocompetent, transformed with the *Mll*Hom5-TAP-FRT-pA-hygro-PGK-FRT-*Mll*Hom3 cassette, and plated on triple selection (cholramphenicol 12,5μg/ml, kanamycin 20 μg/ml and hygromycin 100μg/ml). 15 colonies were analysed by PCR with an analogous strategy to that described for the targeting of the loxP-sA-IRES-βGeok-pA-loxP cassette (see above). The following primers were used: TagScreenF2 and NTagR.

Primer TagScreenF2 has the following 5'-3' sequence:

GGGGTACGGCCATCTGGGCGGCGCCAG

Primer NTagR has the following 5'-3' sequence:

TACAGGCGCGCCGGAAGTTCCTATACTTTCTAGAGAATAGGAACTTCCATCAAG

TGCCCCGGAGGATGAGATTTTC.

14 out of 15 colonies showed an amplification product of the correct size, and their BamHI restriction patterns were analysed through electrophoresis. As shown in figure 38, all colonies displayed the same pattern, indicating that no additional, intramolecular recombination events had occurred. On the basis of these results, clone JC (TAP-*Mll*-LacZ JC) was selected for electroporation in mouse ES cells. DNA was prepared from two liters of bacterial cultures, and then digested with Pi-SceI to release the targeting construct from the BAC vector. Agarose gel electrophoresis confirmed that digestion had occurred to completion, and the size of the actual targeting construct was estimated to be greater than 60 (Fig 39)

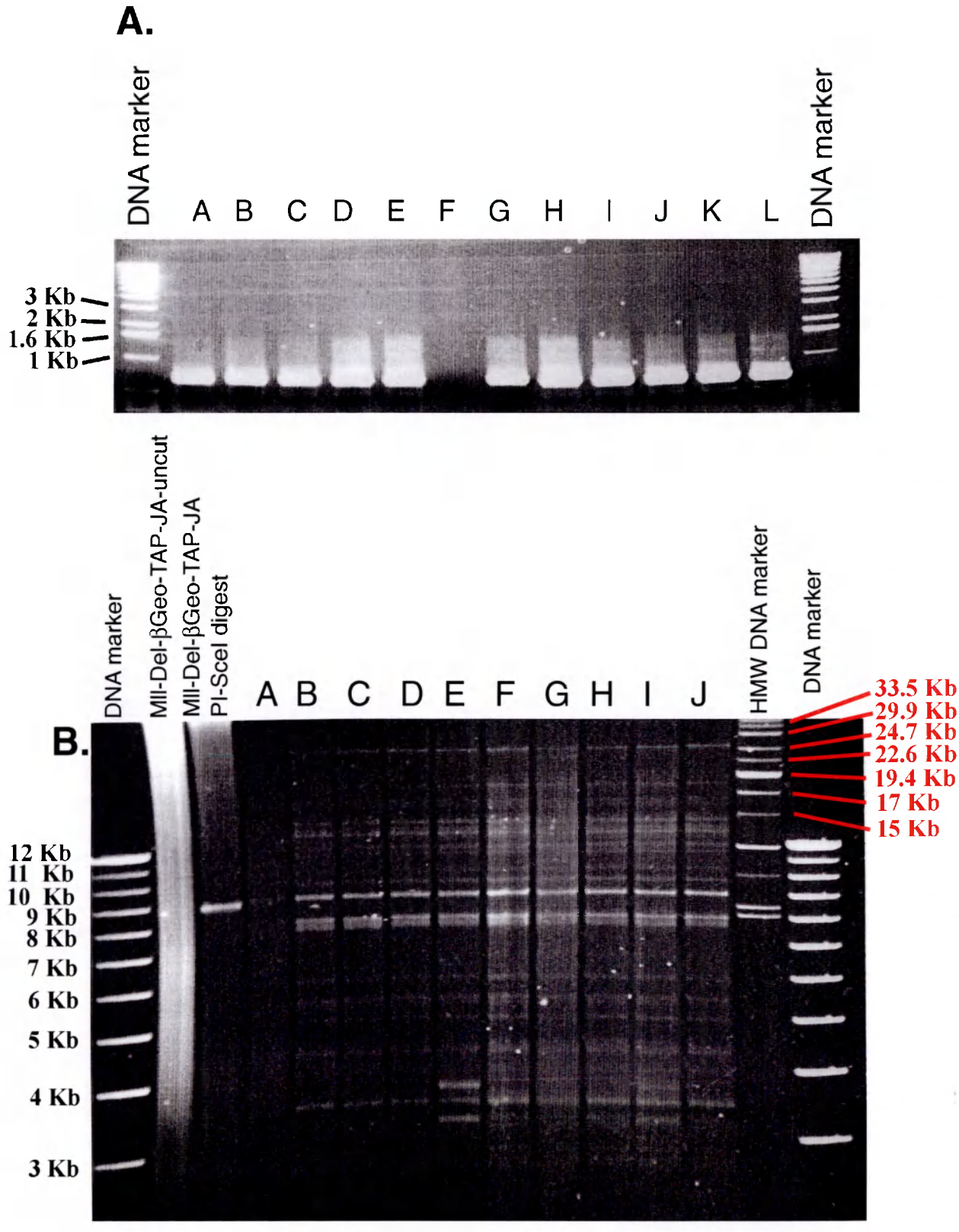


Figure 38 Targeting of the FRT-flanked TAP-Hygro cassette to the shaved MII BAC backbone

The FRT-flanked TAP-hygro cassette was targeted by ET recombination to the first exon of the MII "shaved" BAC backbone already harbouring the LoxP-flanked βGeok cassette in intron 11.

A. PCR screening of candidate colonies with primers spanning the integration site.

B. Represenatative BamHI digest of PCR-positive colonies. All colonies examined showed the same restriction pattern. The two additional bands in colony F result from retention of the recombinogenic plasmid R6K-αβγ.

In a parallel set of experiments, the same knock-in cassette *Mll*Hom5-TAP-FRT-pA-hygro-P-PGK-FRT-*Mll*Hom3 was also targeted to the wild type *Mll* BAC 145L16. The aim was to isolate from this modified BAC a simpler targeting construct which, upon ES cell electroporation, would simply insert the TAP-FRT-pA-hygro-P-PGK-FRT at the ATG of the *Mll* gene. Thereby this conventional targeting construct and the BAC based, longer construct can be compared for targeting efficiencies.

To this end, cells containing both the wt *Mll* BAC 145L16 and the recombinogenic plasmid R6K-αβγ were made electrocompetent and transformed with the *Mll*Hom5-TAP-FRT-pA-hygro-PGK-FRT-*Mll*Hom3 cassette. Cells were plated on double selection (cholramphenicol 12,5μg/ml, and hygromycin 50μg/ml). All 30 colonies picked tested positive by PCR using the same primer pair ScreenTagF2 and NTagR. Colony n. 3 (*Mll*NTag c.3) was sequenced and confirmed to have undergone the correct recombination event. On the basis of available sequence, NheI was selected to digest this targeted BAC and subclone the TAP-tag containing NheI fragment into a high copy vector (in this case pUC34, which carries the ampicillin resistance gene) (Benes et al., 1993) by conventional ligation and double selection for hygromycin and for the selectable marker of the chosen vector. This yielded the second *Mll* targeting vector (hence referred to as *Mll*TAPshort). From colony B, DNA was prepared and digested with NheI to release the targeting construct (*Mll*TAPshort) from the pUC34 backbone, prior to ES cell electroporation.

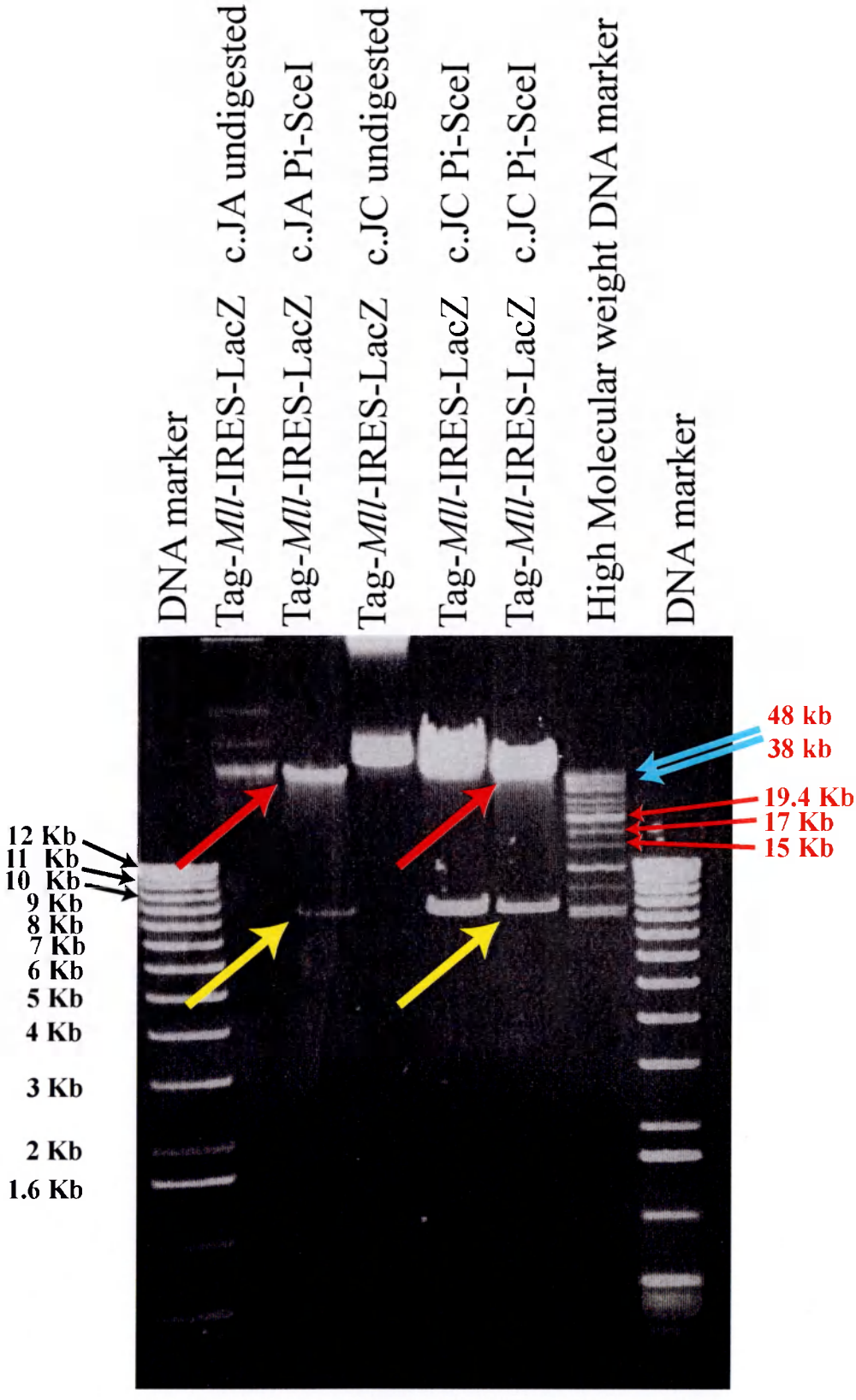


Figure 39 Pi-SceI digests of the *Mll*-TAP-LacZ targeting construct
Agarose gel electrophoresis of Pi-SceI digests from two independent clones of the final Mll targeting construct TAP-*Mll*-LacZ (c. JA and JC). The pattern demonstrates complete digestion. The high molecular weight (HMW) DNA marker permits to approximately estimate the length of the targeting construct (red arrow). The two highest bands of the HMW marker are 48 and 38 kb; thus, the insert fragment is likely longer than 60 Kb. The band representing the BAC backbone (increased by the addition of the two selectable markers introduced in the ET recombination steps) is indicated by a yellow arrow.

IX.4 ES cell targeting with the *Mll* construct.

Mouse ES cells of the E14 line were used for the targeting experiment. Cells were grown on gelatin coated plates in the presence of LIF (see material and methods). A total of 80 μg of the construct TAP-*Mll*-LacZ were digested with Pi-SceI. DNA was phenol-extracted from the digestion mixture, resuspended in (50) μl PBS, and electroporated into 76 * 10^6 cells, divided onto 38 10 cm dishes plus 1 dish for cells which had been electroporated with only PBS as negative control. The plating density was 2,000,000 cells/plate. 24 hours after electroporation, drug selection was started on all plates. Four selection schemes were devised: G418 only (200 μg/ml); hygromycin only (160 μg/ml); G418 (100 μg/ml) plus hygromycin (160 μg/ml); G418 (200 μg/ml) plus hygromycin (160 μg/ml) Several variables impinged on the experiment.

First, there was no conclusive information about the efficiency with which such a big targeting construct would enter ES cells upon electroporation. Very large DNA molecules (mostly in the form of YACs) have been introduced into ES cells by spheroplast or whole cell fusion, a procedure which is likely to yield different results from electroporation.

Second, ES cells are normally transfected and selected using one drug at a time. Previous applications of selection regimens with two different drugs have usually been applied to cells already carrying a first resistance gene stably integrated.

A third variable comes from the configuration of this targeting construct, where hygromycin resistance is under the control of the PGK promoter, while G418 resistance is dependent upon the endogenous *Mll* promoter, whose strength in these experimental conditions was unknown.

Fourth, assuming that the construct would enter ES cells at a workable frequency, a major question concerned the rate at which the construct would break once inside the cells.

Breakage would dissociate the two selectable marker cassettes, since, lying near the ends of the construct, most rupturing would be expected to occur between them. Under such circumstances, G418 resistance would only emerge in those clones where the 3' terminal part of the construct had integrated in a locus permissive for the splicing and expression of the βGeok cassette, whereas hygromycin resistance would emerge from random integration of the 5' end of the construct. By use of either selection alone or in combination, and use of a smaller targeting construct carrying only hygromycin selection, the experiment was designed to address these variables.

IX.4.1 The *Mll*-TAP-LacZ construct integrates mostly as a unit

To investigate the issue of DNA breakage, colonies which had been initially selected with only one drug (either G418 or hygromycin) were replica plated, so that each colony could now be assessed for its resistance to the other drug (percentage of co-resistance).

If the construct lands largely intact in the genome, most of the cells initially selected for hygromycin are expected to be also resistant to G418 (figure 40). In fact, the construct contains the *Mll* promoter and therefore, as long as it stays intact, the activity of this promoter should drive G418 resistance. At the same time, most of the cells initially selected for G418 only are expected to be also resistant to hygromycin, since it carries its own promoter (PGK). Thus, if the construct integrates mostly as a unit, the percentage of co-resistance among these two populations of cells should be very similar.

On the contrary, if the construct breaks (figure 41), the two selectable marker cassettes should behave differently, since the hygromycin cassette carries its own nearby promoter, however breakage will separate the βGeok cassette from its very distant promoter, creating an intron-trap fragment. Cells initially selected for G418 may or may not also be resistant to hygromycin depending on the frequency of unselected random integration

occurring whilst selecting for an intron-trap. Whereas constructs selected for hygromycin resistance only should show relatively lower levels of G418 resistance since this depends on rarer, intron-trap, events.

Two general points need to be made before the results can be interpreted. First, fewer colonies were observed from hygromycin than G418 selection (figure 42). This concords with the greater cytotoxicity and faster killing of hygromycin than GH418. Second, even fewer colonies were observed after double selection, a result indicating that the additional challenge posed by co-selection reduces the number of double resistant colonies.

Figure 40 Integration of the intact *Mll* construct in the genome (diagram)
Schematic representation of the integration of the *Mll* construct in the genome.
If the construct landed in the genome mostly intact, the following prediction could be made: since the construct contains the *Mll* promoter, the hygromycine cassette carries its own promoter while the βGeok cassette is by itself a promoter trap, the majority of clones initially selected in either one of the two drugs should be resistant to both.
Analysis of the percentage of coresistance among colonies initially selected only on one of the two drugs supports the hypothesis that this type of large targeting construct integrates in most cases as a single unit. See text for further details.

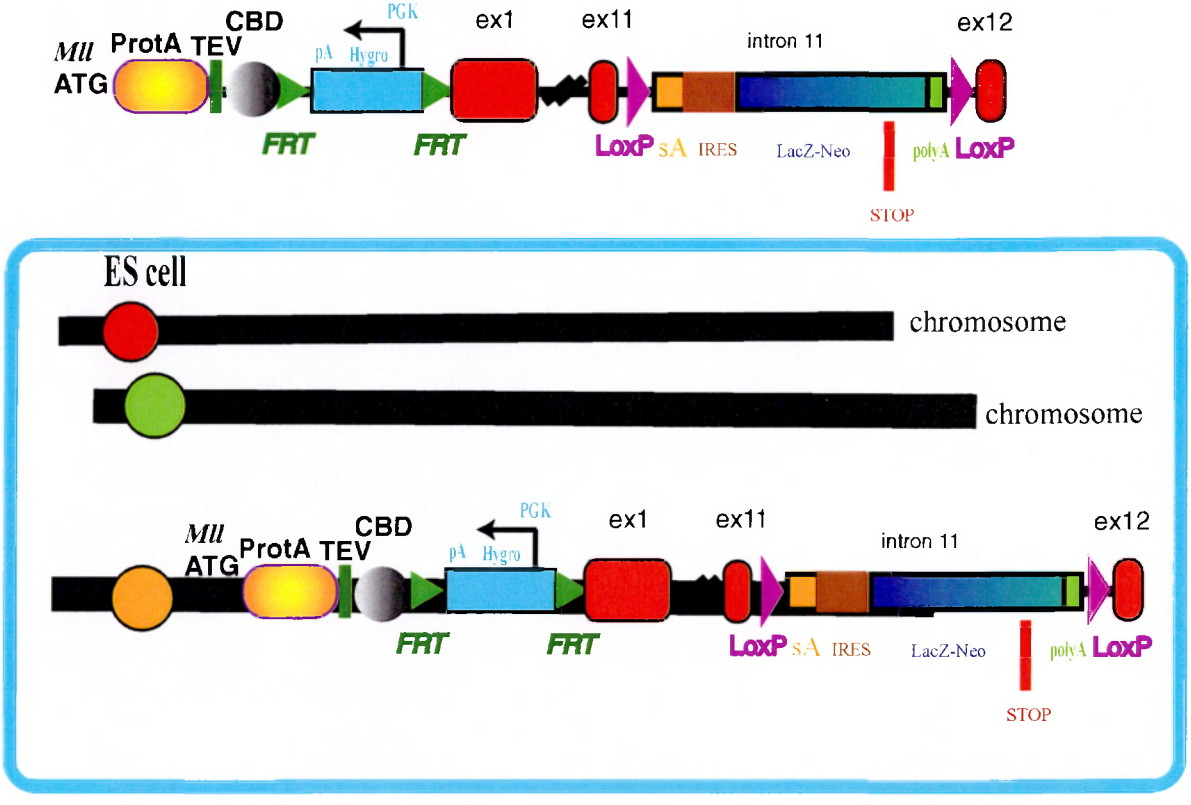
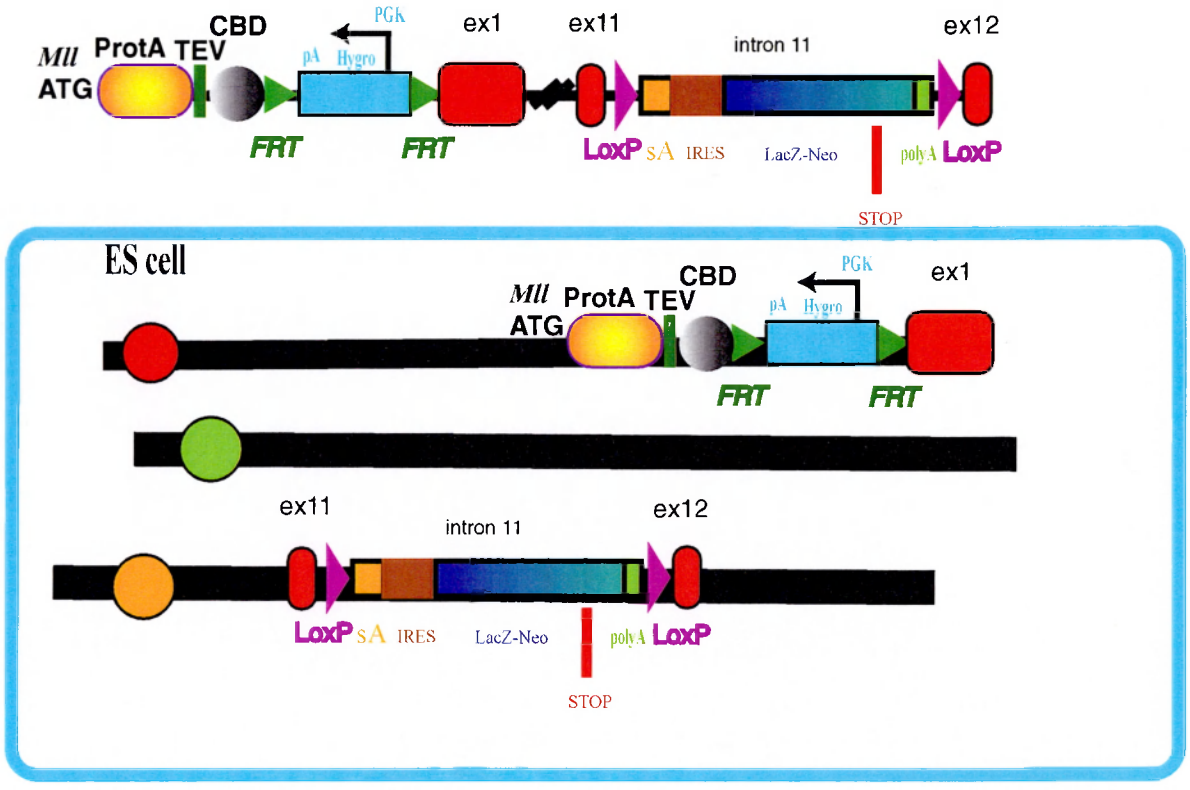


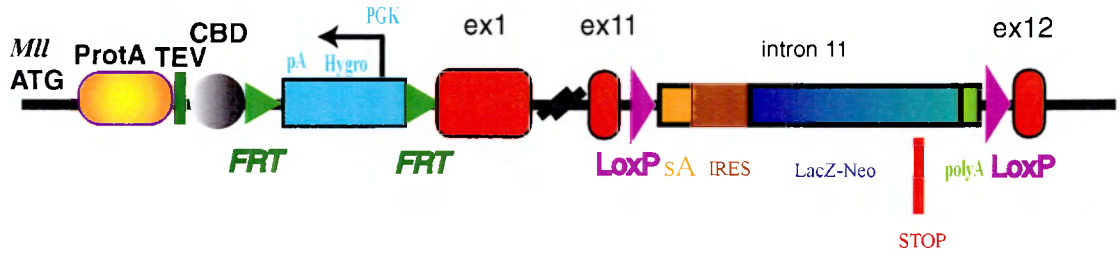
Figure 41 Integration of a cleaved *Mll* construct in the genome (diagram)
Schematic representation of the integration of the *Mll* construct in the genome.
If the construct broke to a great extent before integration, the two selectable cassettes would behave largely independent of each other. Hence the following prediction could be made: the majority of clones selected in G418 should also be resistant to hygromycin, since this cassette carries its own promoter. On the contrary, by selecting only with hygromycin, one could expect a significantly lower number of clones to be resistant also to G418, since this cassette needs to integrate in a permissive spot of the genome to be properly expressed, an event for which there was no selection pressure during treatment with hygromycin alone. Analysis of the percentage of coresistance among colonies initially selected only on one of the two drugs does not support this hypothesis.



In both singly selected cases (figure 42), only a minority (up to 17% for G418 and up to 24% for hygromycin) of resistant colonies were due to construct breakage and integration of only the selected part. On the other hand, the fact that most singly selected colonies were co-resistant and that the total number of doubly selected colonies was about half of singly selected numbers implies that either (i) the construct remains intact upon integration in most cases, or (ii) the frequency of unselected integration after construct breakage is high, or (iii) both (i) and (ii) apply. Since the expected difference between unselected G418 and hygromycin co-resistance frequencies was present but small, the conclusion (ii) is unlikley to account for the majority of events. Hence, conclusion (i) is preferable. It also has the merit of being simpler.

Although not quantitative, these results show that this large construct remains intact during integration in at least a workable frequency of events, and provide evidence that large, BAC based constructs can be used as the platform for the creation of multiply mutagenized alleles in ES cells.

A.



B. 3 different selection schemes

C.

Selection scheme (Micrograms/ml)	G418 200	hygromycin 160	G418 200 hygromycin 160
Colonies/2x10^{6}cells	900	680	450
Co-resistance (%)	83	76	100

Figure 42 Summary of the Mll ES cell targeting experiment
Three different schemes of selection (panel B) were applied to ES cells transfected with the TAP-Mll-LacZ construct (panel A). The results are summarised in panel C. The first row of the table shows the absolute number of colonies/plate obtained in the different selection schemes. The second row indicates the percentage of coresistance to both drugs of colonies initially selected in G418 only (red) or hygromycin only (green). The similarity of the two coresistance indexes (83% vertsus 76%) suggests that the Mll-TAP-LacZ construct integrates in the genome mostly intact.

IX.4.2 Southern hybridisation to identify homologous recombinant clones

The fact that the construct was often incorporated as a unit did not provide per se any information on the frequency of homologous recombinants. Although opening options for random integration strategies based on assembly in E.coli of a single large construct mutated at multiple sites, the application of BACs to simultaneous gene targeting ultimately depends on the rate at which homologous recombinants can be recovered. Previous experiments (Hasty et al., 1991; Thomas et al., 1992) indicated that the efficiency of homologous recombination reached a plateau after about 15 kb of homology arm length. These results may not be directly applicable to the strategy described here, since in the TAP-*Mll*-lacZ construct, the homology arms respectively upstream of the 5' cassette and downstream of the 3' cassette fall within the normal range (5 kb and 6.6 kb respectively). What is different is the presence of about 50 kb of homologous sequence between the two selectable marker cassettes.

In order to identify homologous recombinants, a Southern strategy was devised for both the 5' and the 3' side of the targeting vectors. Both strategies rely on probes which are external to the construct.

BamHI was chosen to explore correct integration at the 5' side. A BamHI site is present 7041 bp upstream of the *Mll* initiating ATG. The next site is 9847 bp downstream. A hybridisation probe was amplified by PCR from *Mll* BAC with primers *Mll*5ProbeF and *Mll*5ProbeR. *Mll*5Probe F has the following 5'-3' sequence:

GGATCAAGAGTAAGCCACATAGCAAGTT

*Mll*5Probe3 has the following 5'-3' sequence:

AATTTGTTGTCCACGGCTTCATCG

The resulting probe is 632 bp long. It detects in all clones a wild type band of 9847 nucleotides. Since there is no BamHI site in the TAP cassette, correct integration of the TAP-*Mll*-LacZ construct results in a higher band of 12176 residues.

XbaI was chosen to explore correct integration at the 3' side. An XbaI site is present in intron 11 8267 bp upstream from the 3' end of the construct. The next site is in exon 19, 1505 bp downstream from the 3' end of the construct. A hybridisation probe was amplified by PCR from *Mll* BAC with primers *Mll*3ProbeF and *Mll*3ProbeR. *Mll*3Probe F has the following 5'-3' sequence:

CTGCCAAACTACTTACGGAAAATG

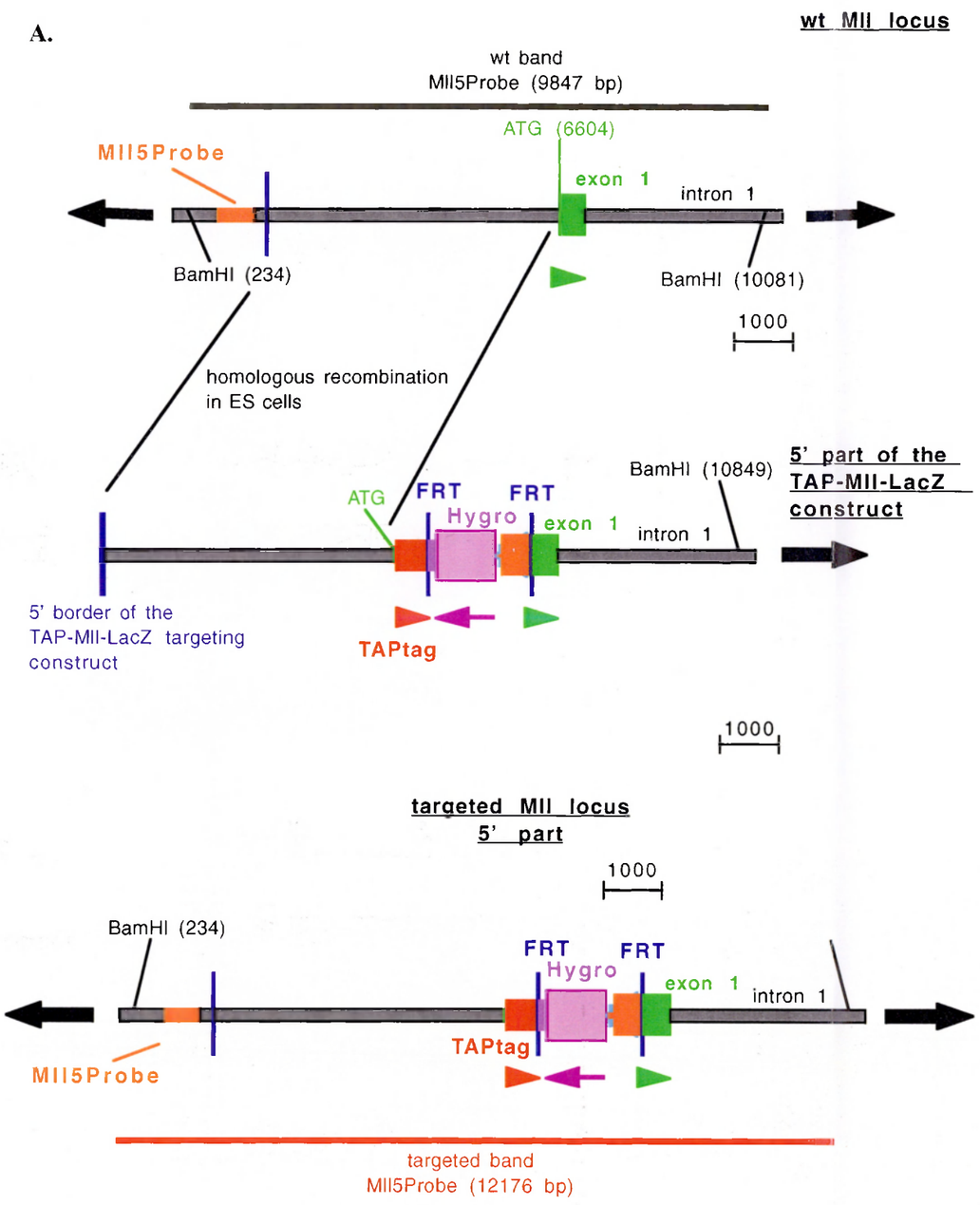
*Mll*3ProbeR has the following 5'-3' sequence:

TTGACACTGAACCACGGAAAAACT

The resulting probe is 858 bp long. It detects in all clones a wild type band of 9773 nucleotides. Since there is an XbaI site in the loxP-sA-IRES-βGeok-pA-loxP cassette, correct integration of the TAP-*Mll*-lacZ construct results in a lower band of 8453 residues.

A total of 29 colonies initially selected in hygromycin (160μg/ml) and G418 (200μg/ml) were analysed by Southern blot for homologous recombination at the 5' end of the locus, using the *Mll*5Probe on BamHI digests. Three colonies (AC4, AD1 and DA4) showed both the band from the wild type allele (9847 bp) and from the homologously recombined one (12176 bp), amounting to a targeting frequency of around 10% (Figure 43).

18 new colonies, again initially selected in hygromycin (160μg/ml) and G418 (200μg/ml), plus 18 colonies already analysed on the 5' side (including the three positives), were then screened by Southern blot with the *Mll*3Probe on XbaI digests to check correct integration at the 3' side of the locus (figure 44). 4 colonies tested positive (AC4, AD1, DA4 and DB2), as shown by the presence of both the wild type band (9773 bp) and the



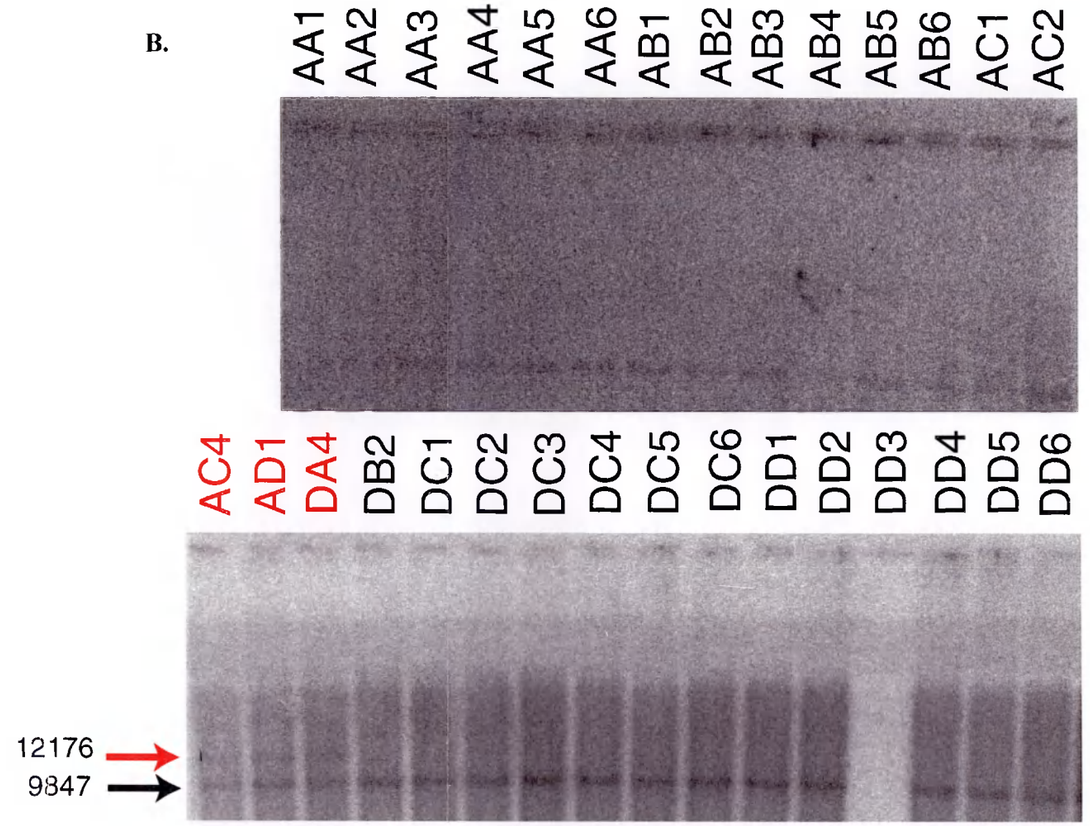
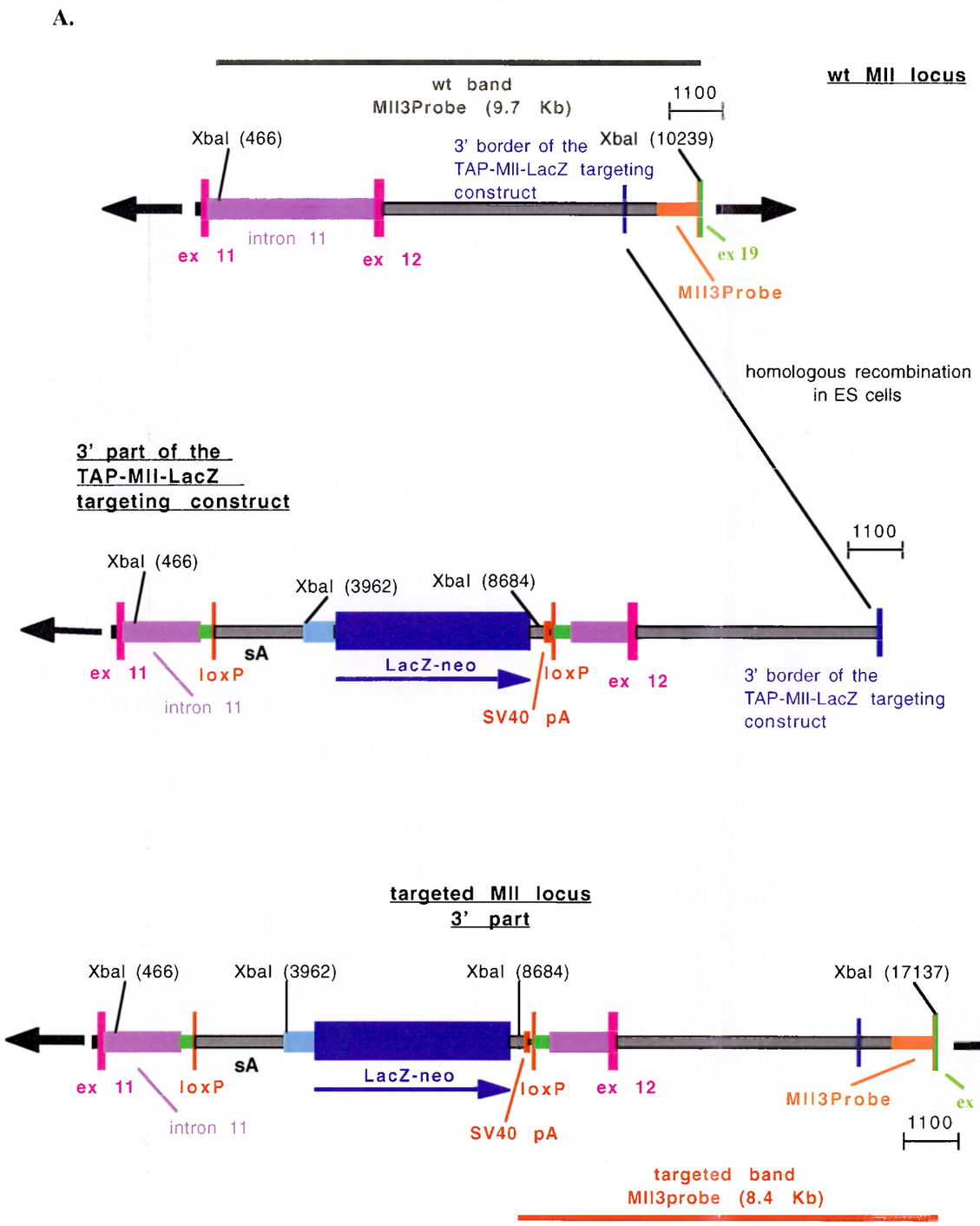


Figure 43 ES cell targeting with the Mll-TAP-LacZ construct (5' side)

A. Schematic representation of the wild type Mll locus, the 5' part of the Mll targeting construct TAP-Mll-LacZ, and the 5' part of the Mll targeted locus. The 5' border of the TAP-Mll-LacZ targeting construct is indicated by a vertical blue line. BamHI sites, on which the Southern strategy was based, are indicated along with their relative position given in parenthesis. Black arrows on either side of the DNA molecules represent flanking genomic regions.
The names of the DNA molecules (constructs and PCR products) are in bold and underlined.
The scale bar indicates the number of nucleotides/centimeter.
Segments of the constructs accompanied by arrows indicate open reading frames.
Abbreviations: (TAP) tandem affinity purification.

B. Southern blot hybridisation with probe Mll5End on BamHI digests of 29 colonies selected in hygromycin (160 mg/ml) and G418 (200 mg/ml) . Colonies AC4, AD1, and DA4 show the correct pattern of wild type and homologously recombined bands (9847 and 12176 bp respectively).



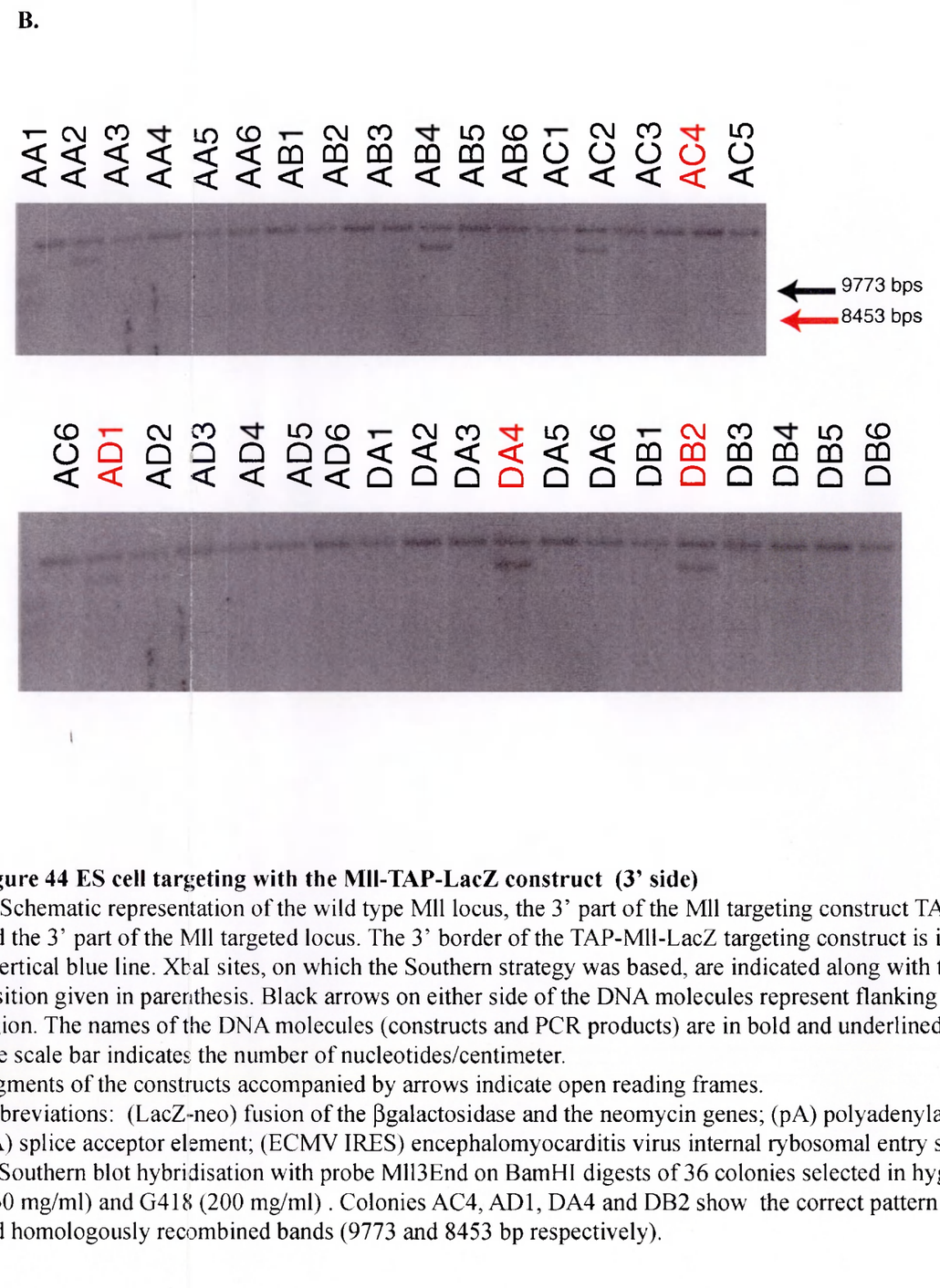


Figure 44 ES cell targeting with the MII-TAP-LacZ construct (3' side)

A. Schematic representation of the wild type MII locus, the 3' part of the MII targeting construct TAP-MII-LacZ, and the 3' part of the MII targeted locus. The 3' border of the TAP-MII-LacZ targeting construct is indicated by a vertical blue line. XbaI sites, on which the Southern strategy was based, are indicated along with their relative position given in parenthesis. Black arrows on either side of the DNA molecules represent flanking genomic region. The names of the DNA molecules (constructs and PCR products) are in bold and underlined.
The scale bar indicates the number of nucleotides/centimeter.
Segments of the constructs accompanied by arrows indicate open reading frames.
Abbreviations: (LacZ-neo) fusion of the βgalactosidase and the neomycin genes; (pA) polyadenylation signal; (sA) splice acceptor element; (ECMV IRES) encephalomyocarditis virus internal rybosomal entry site.

B. Southern blot hybridisation with probe MII3End on BamHI digests of 36 colonies selected in hygromycin (160 mg/ml) and G418 (200 mg/ml) . Colonies AC4, AD1, DA4 and DB2 show the correct pattern of wild type and homologously recombined bands (9773 and 8453 bp respectively).

homologously targeted one (8453 bp). This corresponds to a homologous recombination frequency of about 11%, in good agreement with the results from the 5' side.

3 colonies (AC4, AD1 and DA4) had undergone correct integration at both sides, which, considering the total number of colonies analysed (47) amounts to a targeting frequency of about 6%. This is a very reasonable targeting frequency, similar to, if not higher, than that reported for most targeting experiments. It clearly establishes that homologous recombination in mouse ES cells with a very large, BAC based targeting construct, is practically feasible.

Interestingly, all colonies which tested positive on the 5' side (which contains the hygromycin cassette) had also correctly integrated the 3' end, which harbours the G418 cassette. On the contrary, of the four colonies positive for the G418 cassette, only 3 (75%) had also correctly integrated the hygromycin cassette. Since all these colonies were fully resistant to both drugs, the most likely explanation for this discrepancy is that in one case the construct broke and the hygromycin containing half landed elsewhere in the genome. Although this represents a small sample size, this value (75%) corresponds remarkably with the results obtained from the co-resistance experiment (see above). Taken together, these data suggest that, contrary to previous assumptions, a large DNA molecule can integrate in the genome largely as a single unit. This has very important repercussions for the feasibility of similar strategies. If the construct did undergo substantial breakage with the two mutagenic cassettes integrating independently of each other, the advantage of using a single, big targeting construct over multiple smaller ones would be greatly reduced, as many colonies would probably need to be screened to identify cells which harboured both mutations on the same allele.

IX.5 Further implications of combinatorial gene alleles

The strategy presented here presents new options to combine multiple mutations in a single targeting construct. Ideally, one would want to investigate all mutations independently of each other, which obviously calls for a conditional strategy in which only selected functions of the gene are ablated at a given time in a given tissue, leaving the other domains intact. This can be achieved by flanking relevant domains with site specific recombination target sites, so that upon expression of the corresponding recombinase in a controlled spatio-temporal manner, this domain will be selectively deleted. Currently, two recombinases can be successfully used *in vivo* in the mouse, Cre and Flp. It is clearly a prerequisite for such an approach that other recombinases become available to increase the number of mutations which can be conditionally and simultaneously incorporated in the same mouse line. This is a daunting task, especially since extensive characterisation of the properties of each recombinase is needed for its application as an in vivo genomic engineering tool; however, it is already on the agenda of contemporary mouse genetics.

Thus it is likely that this novel approach to controlled in vivo mutagenesis in the mouse will eventually merge with other large scale projects of contemporary biology and medicine. A framework was recently proposed, along whose lines the sets of data emerging from high-throughput projects (transcriptome, proteome, enzymome, localizome, phenome) can be regarded as functional maps to be incorporated into a multilayered, tridimensional biological atlas (Vidal, 2001). For higher model organisms like the mouse, I propose that an important aspect of this effort will be the generation, in the long term for every gene, of a complete set of conditional alleles (hereafter referred to as the "conditional alleleome") in which every functional domain is flanked by target sites for a different recombinase (Figure 45). In parallel, an exhaustive set of mouse lines should be established, which express different site

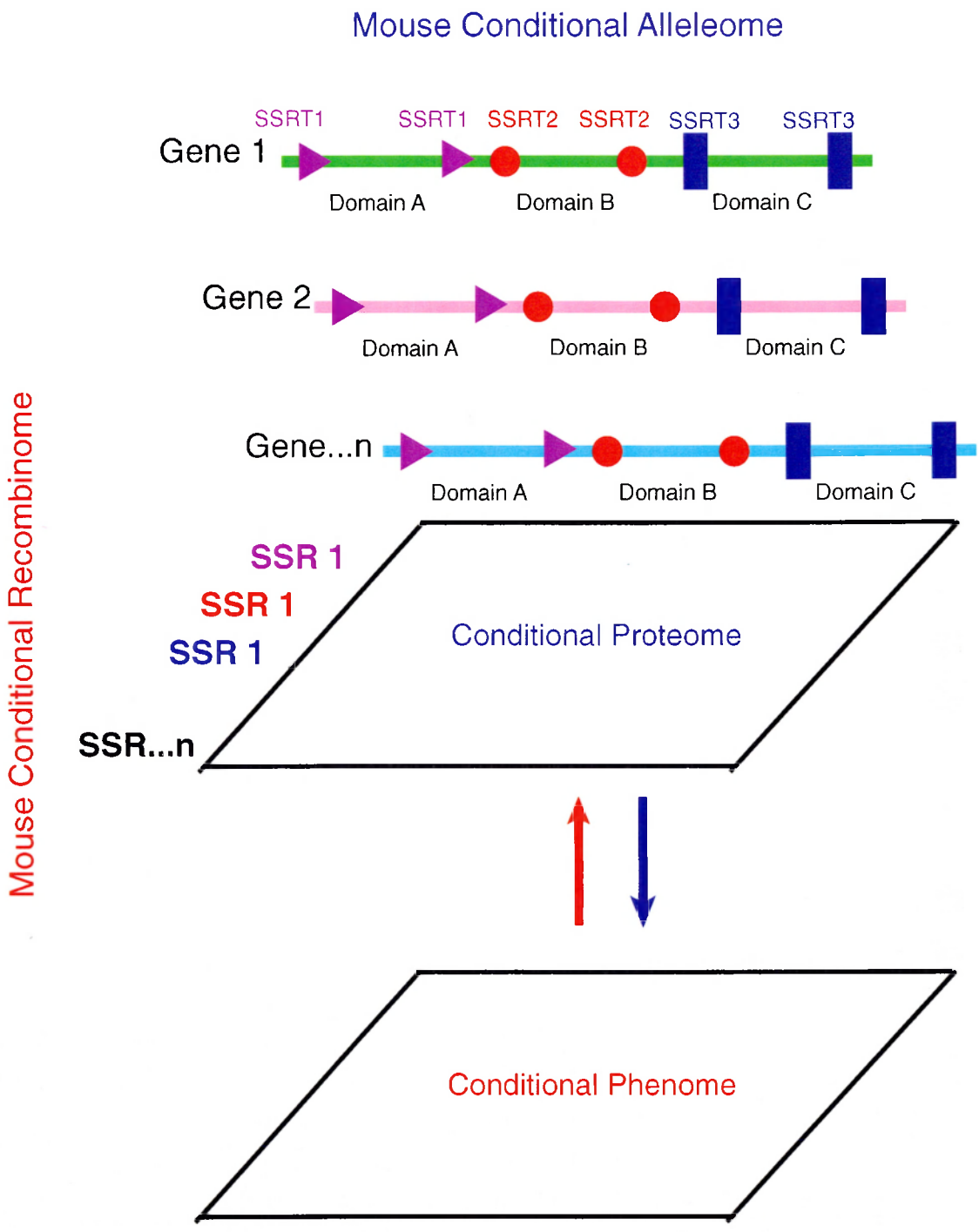


Figure 45 Mouse conditional "alleleome"
Hypothetical diagram of the information generated from a combination of the mouse conditional "alleleome" and "recombinome". On the (x) axis each gene is depicted with its domains flanked by different site specific recombinase target sites (SSRTs), constituting the conditional alleleome. On the (y) axis, mouse lines are represented, each of which expresses a certain site specific recombinase (SSR) in a spatio-temporal controlled manner. Appropriate crossing between these two arrays of mouse lines should enable a proteomic and phenomic analysis centered on the role of individual protein domains.

specific recombinases in different tissues and at different stages of development (hereafter referred to as the "Mouse-Recombinome"). Ideally, they would be conditionally regulatable, most efficiently through fusion with the ligand binding domain (LBD) of nuclear receptors. Multiple crossing schemes combining suitable mouse lines from the "conditional alleleome" and the "mouse recombinome" would then constitute the basis for an accurate functional map of the mouse phenome. If everyone of these alleles includes a protein purification tag, the "conditional proteome" of every gene can be defined, enabling one to pinpoint which domains are responsible for which interaction. Comparison of the results obtained from the "conditional phenome" and "conditional proteome" would yield insight into the (likely) multiple molecular mechanisms underlying a given phenotype, contributing to bridging the gap between the reductionist and organismic interpretation of biological phenomena.

X
Engineering of a Cre mouse line under the control of the *Ikaros* gene

X.1 Overview of the strategy

Studies on leukemogenesis both in human disease and mouse models highlight the importance of the cellular compartment in which fusion proteins are first expressed. As discussed in chapter II.5, the Cre-loxP system offers the potential to explore this central issue. The availability of appropriate Cre expressing mouse lines is therefore crucial, and it is possible to envision that in a not too distant future, the oncogenic potential of the most important translocations could be assessed against a whole blood "Cre zoo" (ie. a collection of mouse lines expressing Cre at key developmental stages for every major hematopoietic lineage).

I decided to establish a transgenic mouse line expressing Cre recombinase under the control of the *Ikaros* gene. The following section highlights some of the key aspects of *Ikaros* function, which provided the rationale for its choice as the gene driving Cre expression.

X.1.1 *Ikaros* in lymphoid development

Ikaros is the founding member of a family of chromatin regulators which play a pivotal role in lymphoid development (Cortes et al., 1999; Georgopoulos et al., 1997; Koipally et al., 1999a). The family includes also Aiolos and Helios (Kelley et al., 1998; Morgan et al., 1997). Its expression is restricted to cells of the hematopietic lineage, with the exception of the striatum in the CNS (Georgopoulos et al., 1992; Morgan et al., 1997). It is present throughout hematopoiesis starting at E8.5 in the mesodermal precursors of the

splanchnopleura which will give rise to all blood lineages, continuing through both yolk sac and fetal liver hematopoiesis, as well as lymphoid development in the bone marrow and in the thymus. In mature blood cells, *Ikaros* is expressed in granulocytes, T, B and NK cells, while it is not expressed in macrophages, mast cells and erythrocytes.

The *Ikaros* gene can give rise, by alternative splicing, to at least eight protein isoforms which mainly differ in the relative content of Krüppel-like zinc fingers, through which Ikaros regulates expression of its target genes and homo or heterodoimerizes with itself or members of the family (Hahm et al., 1994; Molnar and Georgopoulos, 1994). All Ikaros isoforms share the two carboxyterminal Zn fingers required for protein-protein interaction. At the aminoterminus, only some isoforms contain the minimum of two Zn fingers required for binding, with different affinity, to DNA sequences showing variations of the core consensus c/TGGGAAT/c. As multiple copies of Ikaros have been shown to interact in multimeric complexes, the variants which cannot bind DNA are usually referred to as dominant negative isoforms, since their inclusion in such complexes reduces, and eventually abolishes, DNA binding capability (Hahm et al., 1998; Kim et al., 1999; Morgan et al., 1997; Sun et al., 1996).

The relevance of DNA binding for Ikaros function was established by studying loss of function and dominant negative alleles in mice. A null phenotype was generated with a carboxyterminal deletion (Wang et al., 1996). In these mice, B cells and their precursors completely failed to develop, and T cell precursors were also absent throughout fetal development. After birth, however, increasing numbers of thymocytes appeared, although they showed preferential CD4 differentiation and underwent clonal expansions starting from the double positive stage of differentiation. NK cells were also completely lacking, and dendritic cells severely reduced.

Mice homozygous for a dominant negative knock-in allele of *Ikaros*, in which the DNA binding domain is deleted, had an even more severe phenotype, completely lacking also all T cells and succumbing soon after birth to a range of fulminating infections (Georgopoulos et al., 1994). This increased severity is very likely due to a dominant negative effect of this *Ikaros* allele on other members of the family, thus preventing rescue by any possible redundant pathway.

Interestingly, mice heterozygous for this same mutation developed T cell malignancies with 100% penetrance, with eventual loss also of the wild type allele (Winandy et al., 1995).

In the last few years, the molecular mechanism of Ikaros action has been investigated in great detail. Ikaros seems to function both as an activator and a repressor of transcription. It has been found to be associated with at least two distinct multiprotein complexes, the NuRD complex and a Brg-1 based SWI/SNF related complex (Kim et al., 1999). In mature T cells, its relative distribution among these two complexes is approximately 5 to 1. In addition, a small fraction of the total Ikaros protein was found to be associated with Sin3 proteins, which function as transcriptional corepressors (Koipally et al., 1999b). Importantly, these three Ikaros containing complexes have distinct subnuclear distributions which vary with the activation status of the cell. In particular, upon T cell activation, the Ikaros fraction associated with NuRD moves to heterochromatic domains forming toroidal structures visualised by immunofluorescence (Kim et al., 1999). Concomitantly, the SWI/SNF complex shows a diffused pattern. Consistent with these localisation studies, there is a direct correlation between the levels of Ikaros in a mature T cells, and the intensity of signal required for activation; the less Ikaros, the more responsive the cell is to activation, which in turn leads to faster entry and progression through the cell cycle (Avitahl et al., 1999). The

consequence is the accumulation of chromosomal aberrations, likely the molecular explanation for the occurrence of T cell malignancies in heterozygous dominant negative mice.

Recently, several studies have reported the presence of predominantly dominant negative isoforms of Ikaros in infant acute lymphoblastic leukemias, a particularly intriguing finding in view of the acute lymphoid leukemias arising very early in dominant negative *Ikaros* mice (Nakase et al., 2000; Olivero et al., 2000; Sun et al., 1999a; Sun et al., 1999b; Sun et al., 1999c). With variable frequencies among the different leukemias analysed, all these studies demonstrate that blasts from ALL patients express high levels of the dominant negative isoforms Ik-4, Ik-7 and Ik-8, which do not bind DNA and show mainly a cytoplasmic distribution, in agreement with previous results (Sun et al., 1996). On the contrary, normal lymphocytes from a variety of sources (bone marrow, thymus and fetal liver) only expressed Ik-1 and Ik-2 isoforms. This is in agreement with the mouse results, which showed that in normal hematopoetic cells and mature lymphocytes the DNA-binding isoforms make up for the majority of Ikaros proteins expressed.

Of note, in one of these studies, five of five *MLL-AF4+* infants showed this profile of *Ikaros* isoforms expression, and it has been hypothesized that disruption of normal Ikaros function through upregulation of these dominant negative variants could play an important role in ALL, particularly of the infant type (Sun et al., 1999c). Currently, there is no evidence that *MLL* rearrangements are the primary triggering event behind the expression of these splicing variants, mostly because similar ratios of dominant negative vs. full isoforms have been identified also in leukemias characterised by different aberrations. Moreover, the statistics are probably still too small to derive precise correlations. However, the general observation

of an increased level of dominant negative variants holds its validity and appears to be a likely cofactor in leukemogenesis.

Due to its early hematopoietic expression pattern, *Ikaros* is an ideal gene to drive Cre recombinase for loxP modelling of *MLL* leukemias, since they present amplifications of mixed, hence early, lineages. In particular, the leukemic phenotype of *MLL-AF4* is predominantly, but not exclusively, lymphoid, arrested at a pre-B stage of development that also shows concomitant expression of at least one myeloid antigen. A faithful *Ikaros* Cre line should express Cre throughout lymphoid development, including very early stage hematopoietic stem cells, and continuing to mature lymphocytes continously generated in the adult animal.

In the project undertaken here, the suitability of the *Ikaros* expression pattern, through the use of *Ikaros*-Cre mice for induction of the *MLL-AF4* translocation, is pursued. The strategy is a novel extension of the recombinase-steroid hormone ligand binding domain (LBD) fusion protein strategy pioneered in the Stewart lab. It aims to combine the advantage of both Cre-LBDs and Cre-only strategies by use of the Flpe recombinase (Buchholz et al., 1998; Rodriguez et al., 2000). With the use of both Cre-LBD and Cre-only to promote the translocation, it should in principle be possible to directly assess the effect of the translocation on the appearance of the alternatively spliced isoforms of *Ikaros*, for example, amongst other markers.

X.1.2 BAC transgenesis

A BAC transgenesis approach was chosen to express Cre recombinase under the control of the *Ikaros* gene. There are several advantages for using BACs in mouse

transgenesis. The most important one is that, given the large size of the cloned inserts, most or all regulatory regions required for appropriate expression will be contained in the BAC, assuring faithful expression of the desired cassette. Use of a BAC avoids the need to map regulatory regions. Moreover, even if all regulatory elements for a given gene are known, their spatial organisation along the gene locus could be important, thereby preventing cut-and-paste approaches which group regulatory regions within smaller transgenic constructs.

Furthermore, native arrays of regulatory elements in large transgenic constructs are less sensitive to the chromatin organisation at new sites of genome integration. Small transgenic constructs, on the contrary, seem to be sensitive to the integration locus, leading to mosaicism and unpredictable expression patterns. Consequently, a higher number of founder animals need to be analysed in order to identify the appropriately expressing ones.

Another advantage in using BACs or YACs in mouse transgenesis comes form the experience that it is easier to obtain single copy integrants than with smaller, conventional transgenic constructs.

Among the disadvantages of using large genomic clones in transgenesis is the possibility that other genes could also be present in the clone that are dosage sensitive and thereby result in an independent phenotype of their own upon transgenic expression. A related, possibly more dangerous scenario is that truncated genes could be present in the BAC or YAC, coding for proteins which would then act as dominant negative and specifically interfere with various cellular process. While these dangers are real and warrant careful consideration, the experience so far using large genomic clones (BACs or YACs) in mouse trangenesis has shown that they occur very infrequently.

X.1.3 Outline of the "matching pair" Cre-LBD, Cre only strategy

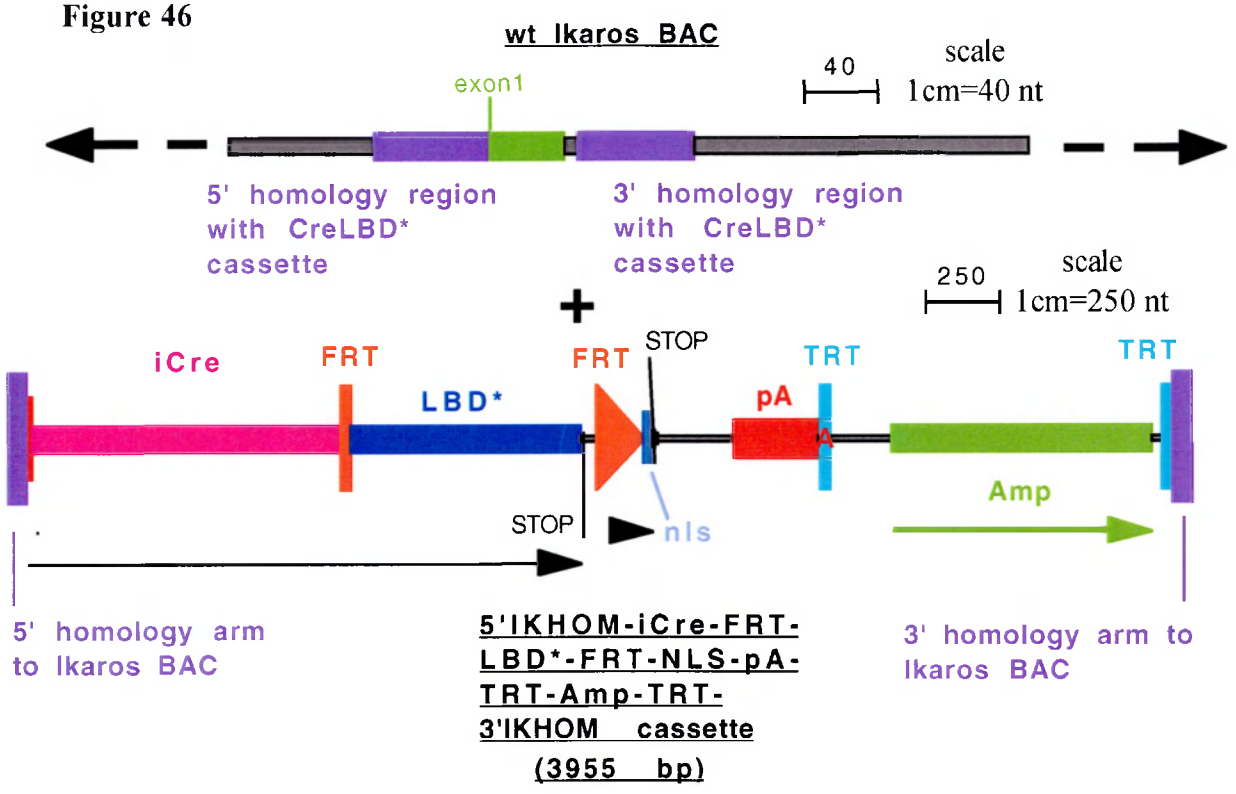
Figure 46 describes the strategy adopted. An improved version of Cre recombinase (in which the codons have been optimized for mammalian usage) was fused to mutated ligand binding domains (LBDs*, where the asterisk indicates the mutation) of either the glucocorticoid receptor (GR) or the estrogen receptor (ER). The fusion of Cre recombinase to nuclear receptor ligand binding domains has been successfully used in diverse experimental settings to regulate Cre mediated recombination through administration of the appropriate ligand, both in cell culture and *in vivo* (see chapter II.5.2).

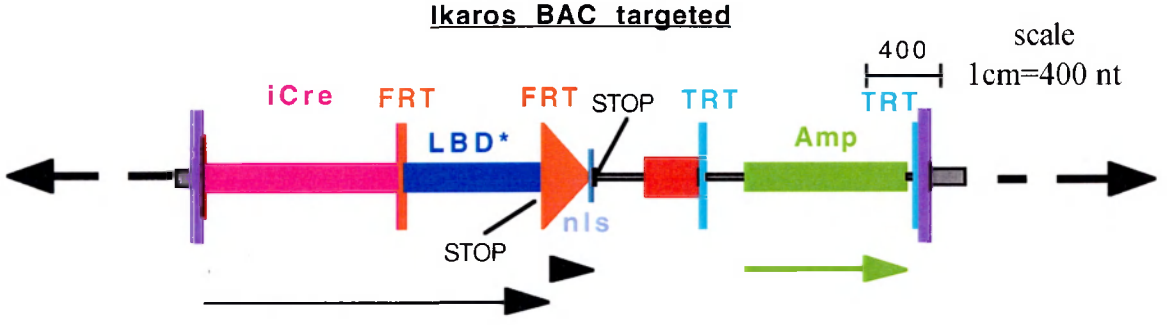
The novelty of the approach outlined here consists in the introduction of two FRT sites which flank the ligand binding domain. The first FRT site is placed in frame between the Cre recombinase and the LBD*. The second FRT is placed immediately after the LBD* upstream of the polyadenylation signal (pA). This FRT is flanked on its 3' side by a nuclear localisation signal, followed by a STOP codon. Upon Flp recombination, the LBD* is deleted, and the resulting cassette should now result in the translation of a constitutively active form of Cre recombinase, fused in frame to an FRT (left after the recombination reaction) and to a nuclear localisation signal for appropriate intracellular targeting. After establishment of transgenic lines carrying an *Ikaros*-Cre-FRT-LBD*-FRT-nls BAC, mice can be crossed to a Flpe deleter to remove the FRT flanked cassette, hence possibly creating a "matching pair" of CreLBD* and Cre-only.

The functional flexibility of this design constitutes a potential improvement in the engineering of Cre expressing mouse lines. In fact, while in some experimental settings, it is highly desirable to have regulated Cre activity, other strategies might require constitutive Cre activity. In the case of the *Mll-Af4* leukemia model, for example, it could be interesting to compare one or multiple pulses of Cre activity (induced by administration of the appropriate ligand) with constitutive Cre activity, with respect to the efficiency of

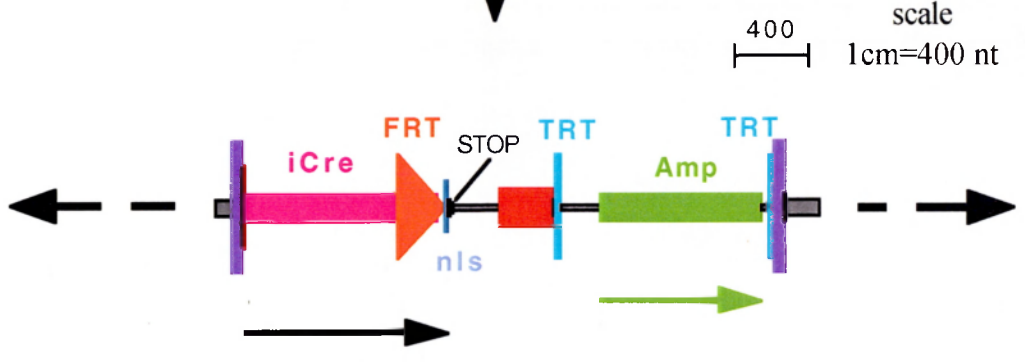
Figure 46 Outline of the "matching pair" Cre-LBD, Cre-only strategy
The Ikaros BAC was targeted with a cassette in which an improved version of Cre recombinase was fused to mutated ligand binding domains of either the glucocorticoid or the estrogen receptor. The ligand binding domains were flanked by FRT sites, resulting, upon Flp recombination, in their deletion and hence in constitutive Cre activity.
The βlactamase gene was used to select for the integration of the cassette into the Ikaros BAC via ET recombination. It was flanked by TRT sites (recognition sites for TnPI recombinase) to allow removal of the selectable marker prior to oocyte microinjection.
The names of the DNA molecules are in bold and underlined. The scale bar indicates the number of nucleotides/centimeter.
Segments of the constructs accompanied by arrows indicate open reading frames.
Abbreviations: (Amp) βlactamase gene; (iCre) improved version of Cre recombinase; (LBD*) ligand binding domain of the glucocorticoid or estrogen receptor;(nls) nuclear localization signal; (FRT) target site for Flp recombinase; (TRT) target site for the TnpI recombinase (pA) polyadenylation signal.
See text for a full description of the cloning strategy.

Figure 46





Flpe recombination



Ikaros BAC targeted after Flpe recombination

interchromosomal translocation. In turn, this could correlate with the frequency of leukemia development, and thus constitute a useful model to address the issue of how many mutated cells are required *in vivo* for cancer development. As virtually all neoplastic diseases in humans are clonal in origin, it is thought that a single cell needs to be hit by the oncogenic event(s) for the neoplasm to develop. However it remains possible that within an organism, multiple cells need to undergo a certain genetic lesion before one clone effectively manages to establish a tumor.

X.2 Assembly of the *Ikaros*-Cre BAC transgenes

X.2.1 Assembly of the CreFRTGBD*FRT construct

As a preliminary step, the core components of this conditional-constitutive Cre cassette were assembled in the construct pBKC-iCreFRTGBD*FRTNLS.

The starting point was the construct pBKC-iCreFRTGBD*, which had been previously generated (Michael Hübner, unpublished results). The FRT-NLS-STOP module was cloned in pBKC-iCreFRTGBD* by KpnI-PflMI ligation. For this, the following oligonucleotides were designed: FRTNLSupper and FRTNLSlower.

Oligonucleotide FRTNLSupper has the following 5'-3' sequence:

TTGGACTAGAGGTACC**GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTC**ACC AAAGAAGAAGCGAAAGGTC*TGA*CACAGTAC.

The FRT (in bold) is located at positions 17-50. A single nucleotide (dATP, in bold and underlined) is inserted immediately downstream of it to maintain the reading frame with the nuclear localisation signal (PKKKRKV), coded by residues 52-72 (underlined). A STOP codon is inserted at position 73-75 (italics). The first four residues (TTGG at position 1-4) constitute the upper strand of the PflMI 3' overhang for cloning into pBKC-iCreFRTGBD*. The last five residues (AGTAC at position 79-83) constitute the upper strand of the KpnI 5' overhang for cloning into pBKC-iCreFRTGBD*. The A at position 79, which replaces the G, was inserted to destroy the KpnI site to permit subsequent cloning steps.

Oligonucleotide FRTNLSlower has the following 5'-3' sequence:

TGTGTCAGACCTTTCGCTTCTTCTTTGGTGAAGTTCCTATACTTTCTAGAGAATAG GAACTTCGGTACCTCTAGTCCAAGTC.

It is complementary to the sequence of FRTNLSupper. At the 5' and 3' ends, it creates the appropriate lower strand overhangs for cloning into PflMI and KpnI sites respectively.

The two primers FRTNLSupper and FRTNLSlower were annealed to each other and ligated into pBKC-iCreFRTGBD*, to yield pBKC-iCreFRTGBD*FRTNLS-STOP. 8 colonies were analysed by EcoRI and XbaI digestion and showed the correct pattern. Colony n.8 (pBKC-iCreFRTGBD*FRTNLS-STOP c8) was sequenced to check the integrity of both the Cre cassette and the FRTNLS-STOP module and selected for the successive cloning steps.

The next step involved ET mediated insertion of the Cre cassette under the control of the *Ikaros* regulatory regions present in the two different *Ikaros* BACs.

Both *Ikaros* BACs (see above) contained the initiating ATG of the *Ikaros* gene. A full contig of 11409 bp had been sequenced previously and found to be present in both BACs (data from Dr. Meinrad Busslinger). It includes 6729 bp upstream of the initiating ATG, and 4614 bp downstream of it. The ET cloning strategy involved placing the Cre cassette immediately downstream of the *Ikaros* ATG, so that it would be transcribed under the full array of regulatory elements of the *Ikaros* gene. Therefore, the 5' homology arm was designed so that it would encompass 61 nucleotides upstream of the ATG. As for the 3' homology arm, it would have been possible to leave intact the rest of the first exon by targeting the cassette between the first and the second codon of exon 1. However, in this configuration, splicing could have taken place between exon 1, now harbouring the Cre cassette, and the rest of the gene. Although the presence of a polyadenylation signal within the Cre cassette would theoretically stop transcription, aberrant splicing could have resulted in the production of a chimeric protein containing portions of the CreLBD* cassette upstream of the Ikaros domains, which might have interfered with normal function. In order to reduce the possibility of aberrant splicing, the 3' homology arm was designed to include 61 residues starting 6 bp after the start of intron 1. Homologous recombination would thus result in the deletion of the first exon and the first six residues of intron 1.

To place the Cre cassette immediately downstream of the *Ikaros* ATG, PCR was avoided by cloning the 5' ET homology arm between the EcoRI and HindIII sites located upstream of the starting ATG of the Cre gene. The original *Ikaros* sequence upstream of the ATG was left intact, immediately followed by the twelve nucleotides downstream of the HindIII sites in pBKC-iCreFRTGBD*FRTNLS-STOP which precede the starting ATG of the Cre gene. The insertion of twelve nucleotides between the whole *Ikaros* regulatory region and the initiating ATG of Cre should not have any impact on transcription or translation of the Cre protein. Furthermore, in terms of translation efficiency, the sequence immediately upstream of the ATG is a better match to the Kozak consensus sequence than the original *Ikaros* one. Whereas the original *Ikaros* sequence reads 5'GAAGACA 3', the Cre ATG is preceded by the sequence 5' GTCCACC 3', where ACC is a perfect match for the Kozak consensus (Kozak, 1986; Kozak, 1987).

To clone the 5' ET homology arm into pBKC-iCreFRTGBD*FRTNLS-STOP, the following oligonucleotides were designed: IKHOMF and IKHOMR.

Oligonucleotide IKHOMF has the following 5'-3' sequence:

AATTC*GCGGCCGC***CATATTTTTGGTTTAAAGTAAAATCCATTTCTCTCTTCTC TTCTCAGATAACCTGAAGACA**A.

Residues 13 through 74 constitute the arm of homology to the 61 residues upstream of the intiating ATG of the *Ikaros* gene. A NotI site (positions 6-12, italics) was inserted to release the recombinogenic cassette iCreFRTGBD*FRTNLS-STOP from the pBKC vector. The first five residues (AATTC) and the last residue (A) constitute the upper strand overhangs for cloning respectively into the EcoRI and HindIII sites of pBKC-iCreFRTGBD*FRTNLS-STOP.

Oligonucleotide IKHOMR has the following 5'-3' sequence:

AGCTTTGTCTTCAGGTTATCTGAGAAGAGAAGAGAGAAATGGATTTTACTTTAA ACCAAAAATATGGCGGCCGCG.

It is complementary to the sequence of IKHOMF. At the 5' and 3' ends, it creates the appropriate lower strand overhangs for cloning into EcoRI and HindIII sites respectively. The two oligonucleotides were annealed and ligated into pBKC-iCreFRTGBD*FRTNLS-STOP c8. Six colonies were analysed through DraI digestion and showed the correct pattern. Colony n.1 (pBKC-5'IKHOM-iCreFRTGBD*FRTNLS-STOP c1) was selected for the subsequent rounds of cloning.

In order to be able to select for the ET mediated recombination of the Cre cassette into the *Ikaros* BAC, a resistance marker gene was needed. The β-lactamase gene which confers resistance to ampicillin was chosen and placed immediately upstream of the 3' ET homology arm. However, selectable marker cassettes can affect the expression pattern of neighbouring or targeted genes (Fiering et al., 1993; Kim et al., 1992; Olson et al., 1996), and long-range interferences at distances greater than 100 kb have been reported (Pham et al., 1996). Furthermore, the neo resistance gene contains cryptic splice acceptor sites (Carmeliet et al., 1996; Meyers et al., 1998; Nagy et al., 1998). Although in the configuration of the Cre cassette the selectable marker would lie downstream of the polyadenylation signal, secondary effects on transcriptional regulation could not be excluded. Therefore, it was desirable to remove the selectable marker from the modified BAC after ET recombination, so that the final targeted BAC injected into mouse oocytes would only contain the Cre cassette. Usually, removal of a selectable marker is achieved by flanking it with recognition sites for a site specific recombinase (like Cre or Flp), which then mediate its excision. However, the design of the Cre cassette described above prevented utilisation of either Cre or Flp. In fact, the presence of the FRT sites flanking the LBD* ruled out Flp as a possibility. LoxP sites could have been used, but upon recombination one

would have been left in the targeted BAC and eventually inserted in the mouse genome. This could cause undesired chromosomal aberrations when the *Ikaros*-Cre mouse line would be crossed to any mouse line already harbouring loxP sites in the genome.

This difficulty highlights the increased need to develop more recombinases as tools in genome engineering. This is of immediate importance for DNA engineering in E.coli, where, as exemplified here, the usefulness of Cre and Flp was already saturated. In the mouse, the recombinational power of Cre and Flp is at present far from its limit, but it is conceivable that in a not too distant future, strategies like the one described for the multifunctional *Mll* allele will become widespread. As these strategies rely on the availability of multiple SSRs target sites (SSRTs) to simultaneously flank different portions of a gene, the need for additional SSRs active in mammalian organisms will also become apparent.

In the case of the *Ikaros* Cre cassette, the problem was solved by using the TnpI recombinase, thereby providing evidence for the usefulness of this recombinase in BAC engineering (see section X.2.5). A PCR strategy was developed which incorporated in the oligonucleotides both the 3' ET homology arm, the newly defined TnpI recognition sites (TRTs) flanking the β-lactamase gene, and the portion annealing to the β-lactamase template.

The following oligos were designed: TRTAmpF and TRTAmpR.

Oligo TRTAmpF has the following 5'-3' sequence:

TTAGGTTC*ACGCGT***TAATACAACACAATATTAATTGTGTTGTA***TTAATTAA*TGA AGACGAAAGGGCCTCGTGATACGCC

The TRT site (in bold) is located at position 15-46. A PacI site was inserted (position 43-50, italic) to facilitate identification of colonies having undergone ET recombination. The Mlu I site (residues 9-14, italic) was placed to enable cloning of the PCR product into the MluI site

of pBKC-5'IKHOM-iCreFRTGBD*FRTNLS-STOP. As a source of template for the β-lactamase gene, the pACYC177 vector was used. The portion of this oligo which anneals to pACYC177 includes residues 52-79 (underlined).

Oligo TRTAmpR has the following 5'-3' sequence:

TGCAA*CGACGCGT*GCGGCC**GCCTTGCTAGCTTCCCCAAGGACAGCTAGCACA ACTTCTGGACATACAACTATCACCCAAGG**<u>**TAATACAACACAATTAATATTGT GTTGTATTAAT**CAATCTAAAGTATATATGAGTAAACTTG</u>.

Residues 20 through 81 (in bold) comprise the arm of homology to the region of the *Ikaros* BAC immediately downstream of exon 1 (corresponding to residues 6838 to 6899 of the *Ikaros* contig). The TRT site is located at positions 82-113 (in bold and underlined). The Mlu I site (residues 8-13, italic) was placed to enable cloning of the PCR product into the MluI site of pBKC-5'IKHOM-iCreFRTGBD*FRTNLS-STOP. A NotI site (positions 6-12, italic) was inserted as an anchor point to release the recombinogenic cassette iCreFRTGBD*FRTNLS-STOP from the pBKC vector. The portion of this oligo which anneals to pACYC177 includes residues 114-143 (underlined).

The resulting PCR product (TRTAmpTRTIKHOM3) was cloned into the MluI site of BKC-5'IKHOM-iCreFRTGBD*FRTNLS-STOP c1. Ten colonies were analysed through XmnI digestion and half had integrated the PCR product in the correct orientation. Colony 6 was sequenced to check the integrity of the TRT sites and selected for ET recombination into the two *Ikaros* BACs.

X.2.2 Assembly of the CreFRTEBD*FRT construct

As an alternative to the 5IKHOM-iCreFRTGBD*FRTNLS-TRTAmpTRT3IKHOM, another construct was assembled in which the improved Cre recombinase was fused to the

triple mutated version of the estrogen receptor ligand binding domain (EBD*)(Feil et al., 1997; Indra et al., 1999). The cloning strategy involved PCR amplification of the triple mutant EBD* module followed by subcloning into pBKC-5'IKHOM-iCreFRTGBD*FRTNLS-STOP-TRTAmpTRTIKHOM3-c6 to exactly replace the GBD* with the EBD*. The triple mutated EBD* module was amplified by PCR (template was a gift of Dr. Gunther Schütz) using the following oligos: NewEBDF and NewEBDR.

Oligo NewEBDF has the following 5'-3' sequence:

TGGGGA*CGTACG*GG**GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTC**TGCCA ACCTTTGGCCAAGCCCGCTCATG.

A BsiWI site was inserted at position 7-12 (italic) to allow directional cloning into the corresponding BsiWI site present in the 5IKHOM-iCreFRTGBD*FRTNLS-TRTAmpTRT3IKHOM. The FRT was inserted in the oligo (positions 15-48, in bold), immediately followed by the portion of the oligo which anneals to the EBD* containing template (residues 49-76, underlined).

Oligo NewEBDR has the following 5'-3' sequence:

AACTTC*GGTACC*TCTAGTCCAATCAGACTGTGGCAGGGAAACCCTCTGC.

A KpnI site was inserted at position 7-12 (italic) to allow directional cloning into the corresponding KpnI site present in the 5IKHOM-iCreFRTGBD*FRTNLS-TRTAmpTRT3IKHOM. The portion of this oligo annealing to the EBD* containing template comprises residues 23 through 49 (underlined).

The resulting PCR product (NewEBD*) was subcloned into the BsiWI and KpnI sites of 5IKHOM-iCreFRTGBD*FRTNLS-TRTAmpTRT3IKHOM. Recombinant colonies were analysed by HindIII digestion. Positive colonies were sequenced to exclude the presence of mutations in the EBD* introduced during PCR amplification and Colony n.5 (5IKHOM-

iCreFRTEBD*FRTNLS-TRTAmpTRT3IKHOM c5) was selected for ET recombination into both *Ikaros* BACs.

X.2.3 ET mediated targeting of the Cre cassettes into the *Ikaros* BACs

Two different *Ikaros* BACs (n.6 and n.3) were obtained (Dr. Meinrad Busslinger, unpublished results). As a preliminary step, both BACs were analysed with a panel of restriction digests (PacI, BamHI, NotI, SpeI, SbfI). Figure 47 shows that the two BACs share as expected some common bands, but overall they have clearly different patterns. Since no information was available about the respective amount of *Ikaros* regulatory regions contained in each BAC, it was decided to target both BACs with the Cre-LBD* cassettes.

Prior to injection into fertlised mouse oocytes, both BACs had to undergo three rounds of modification.

First, the CreLBD* cassette was inserted into exon 1 of the *Ikaros* gene in the first ET mediated recombination step.

Second, the loxP site present in the BAC backbone had to be eliminated to avoid undesired chromosomal rearrangements when the *Ikaros*-Cre mouse line would be crossed to any mouse line harbouring loxP sites. There would be two approaches to address this problem. The targeted BAC insert could be released from the BAC backbone prior to injection, for example by digestion with the NotI sites also present in the BAC vector, or the loxP site could be deleted through ET recombination. As shown in figure 47, it seemed likely that a NotI site could also be present in the BAC inserts of the two *Ikaros* BACs, thus preventing their use for linearisation. Moreover, although injection of a linearised transgene is the standard procedure for normal-size constructs, in the case of BACs the use of a linearised construct increases viscosity (thus making injection more difficult and less controllable) and reduces the average number of transgenic founders. Plus, it would be

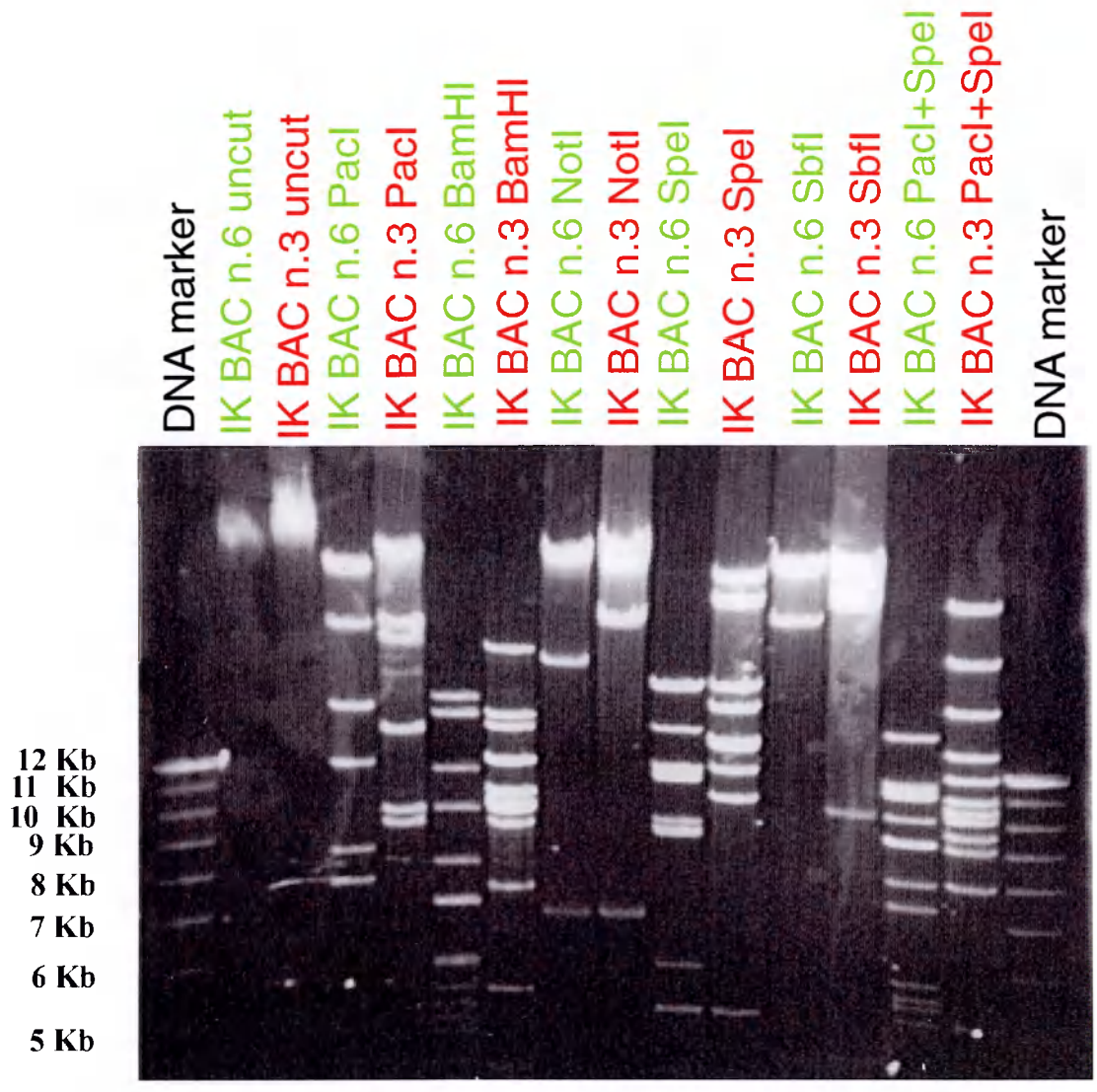


Figure 47 Restriction digests of Ikaros BACs
0.4% agarose gel electrophoresis of multiple restriction digests from two independent Ikaros BACs as a preliminary step for their characterisation and subsequent manipulation via ET recombination.

valuable to have a transgenic constuct which could be injected both in a supercoiled or linearised format to compare the results. Thus, the two *Ikaros* BACs, after having been targeted with the Cre cassettes, underwent a second round of ET recombination to replace the loxP site with a selectable marker.

Third, cells carrying the doubly targeted BACs were transformed with a plasmid expressing the TnpI recombinase, in order to delete the β-lactamase gene from the Cre cassette.

For the first ET recombination step, HS996 cells harbouring either of the two *Ikaros* BACs were transformed with the recombinogenic plasmid R6K-αβγ. Electrocompetent cells were prepared from one colony for each BAC (Colony E for BAC n.6 and colony H for BAC n.3), electroporated with the recombinogenic cassette 5IKHOM-iCreFRTGBD*FRTNLS-TRTAmpTRT3IKHOM and plated under double selection (chloramphenicol 12,5 μg/ml and ampicillin 50 μg/ml).

Recombinant colonies were preliminarily identified by PCR screening with the following primers: *Ikaros*2F and IKProbe1R. These primers yield an amplification product only if the Cre cassette has been integrated in the correct site in the *Ikaros* BACs.

Primer *Ikaros*2F has the following 5'-3' sequence:

ATTCACAGTCCCAAGGCTCATTTC.

It anneals to the beginning of the PolyA signal in the Cre cassette cassette.

Primer IKProbe1R has the following 5'-3' sequence:

AAAATGCAAACTTACAAACAAATCAA.

It anneals to the *Ikaros* contig 1645 residues downstream of the 3' ET homology arm. The resulting PCR product is 2193 bp long.

10 colonies were analysed for each *Ikaros* BAC, and all tested positive for the presence of the correct amplification product (Figure 48). 4 colonies (n. 1 and n. 10 from the BAC 3H, and n. 3 and n.8 from BAC 6E) were sequenced to check the correctness of the integration point, and all yielded the correct result. After checking that they still contained the recombinogenic plasmid R6K-αβγ, competent cells were prepared from all four colonies for the subsequent ET recombination step.

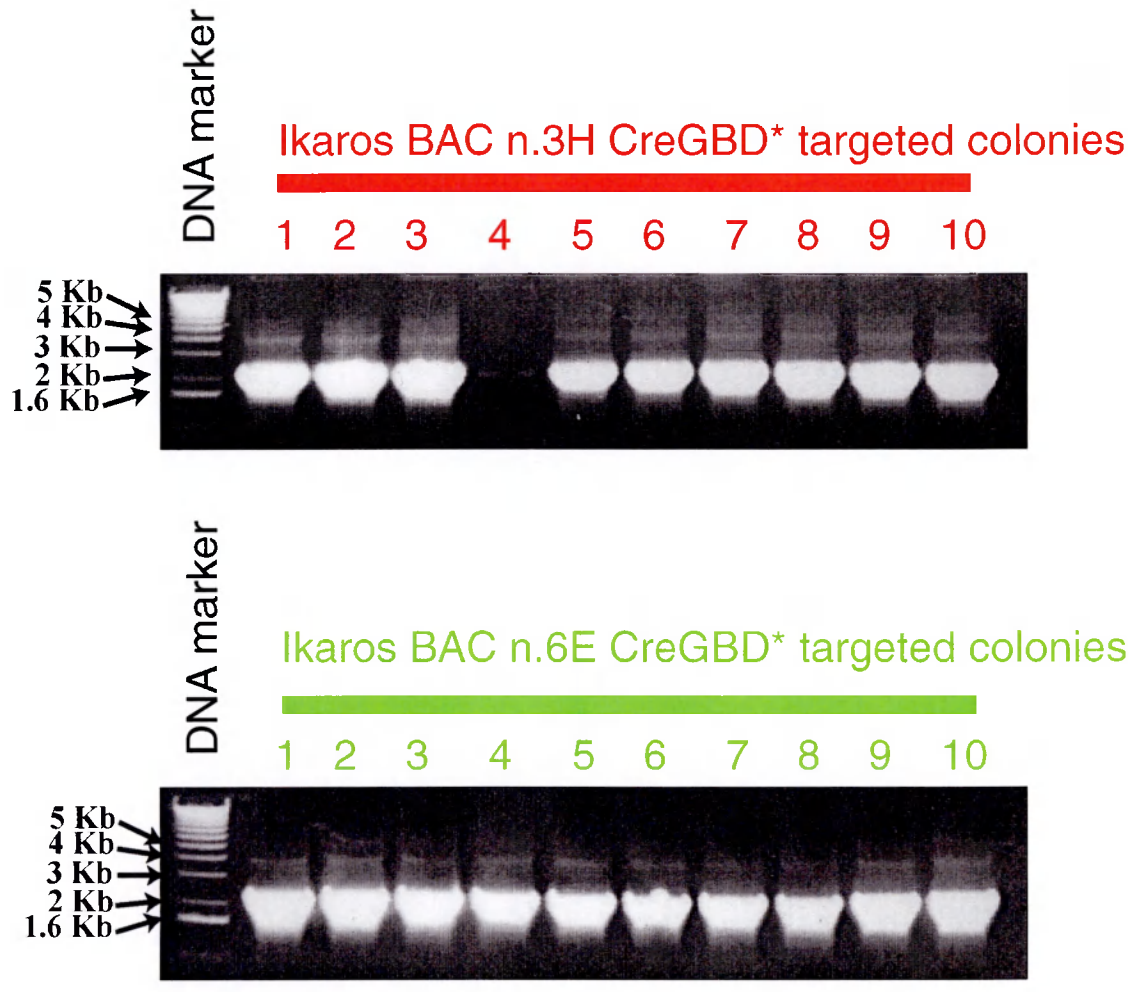


Figure 48 PCR screening of Ikaros BACs targeted with the CreGBD* cassette
Analysis of 20 candidate colonies after ET mediated insertion of the CreGBD* cassette in the two Ikaros BACs n.3 and n.6. PCR amplification with primers Ikaros2F and IKProbe1R yields a product of 2193 bps.

X.2.4 ET mediated deletion of the BAC vector loxP site

To delete the loxP site present in the BAC backbone, the gentamycin resistance gene was chosen as a selectable marker. Oligos were designed to amplify a product where the gentamycin resistance gene was flanked on both sides by homology arms to direct its integration into the BAC backbone. Oligo *Mll*Del2R has been described above for the second deletion step in the assembly of the *Mll* construct. Residues 1-65 constitute the homology arm to the BAC backbone. The portion of the oligo annealing to the gentamycin resistance gene (from plasmid pBADαβγ-Gentamycin) comprises residues 66-90.

Oligo BACgentF has the following sequence 5'-3':

GCCCCGACACCCGCCAACACCCGCTGACGCGAACCCCTTGCGGCCGCATCG

AATGCCCGGGCTAGGGATAACAGGGTAATCTATGTCGGGTGCGGAGAAAGAGG

TAATGAAATGGC*TGAAGGCACGAACCCAGTTGACATAAGCC.*

Residues 1 through 54 (in bold) constitute the 5' arm of homology to the BAC backbone, immediately upstream of the loxP site. A rare cutter polylinker (positions 55-116, underlined) was inserted to permit eventual linearisation of the targeting cosntruct prior to oocyte microinjection. It features the following restriction enzyme recognition sites: SrfI (55-62), I-SceI (63-80) and Pi-SceI (79-116). The portion of the oligo which anneals to the Gentamycin resistance gene template includes residues 117-145 (italics).

The resulting PCR product (BACGent) was electroporated into competent cells prepared from the four colonies originating from the previous ET cloning step (n. 1 and n. 10 from the BAC 3H, and n. 3 and n.8 from BAC 6E). Cells were plated on triple selection (chloramphenicol 12,5 μg/ml, ampicillin 50 μg/ml and gentamycin 3μg/ml).

Recombinant colonies were preliminarily identified through a PCR screening with the following primers: GentF and BACRecR. These primers yield an amplification product only if the gentamycin cassette has been integrated in the correct site in the BAC backbone.

Primer GentF has the following 5'-3' sequence:

ACGATGTTACGCAGCAGGGCAGTC.

Primer BACRecR has the following 5'-3' sequence:

CGTCGGTCTGATTATTAGTCTGGA.

The resulting amplification product is 1384 bp long. Sequence analysis was initially performed on representative PCR positive colonies originating from BAC n.6 (from both batches of competent cells n. 3 and n. 8). Colony O was found to be correct (IK6E-CreGBD-Gent-C.3/O) and was chosen as such for injecting into mouse oocytes. This targeted BAC still harboured the β-lactamase gene downstream of the Cre cassette.

In a parallel set of ET targeting experiments, the conditional CreEBD* cassette was placed into both *Ikaros* BACs, with the same procedure described for the conditional CreGBD cassette. The recombinogenic cassette 5IKHOM-iCreFRTEBD*FRTNLS-TRTAmpTRT3IKHOM was electroporated into the *Ikaros* BAC competent cells (Colony E for BAC n.6 and colony H for BAC n.3), and cells were plated on double selection (chloramphenicol 12,5 μg/ml and ampicillin 50 μg/ml).

To identify correct recombinants, the same PCR screening strategy described above was applied, which relies on primers *Ikaros*2F and IKProbe1R. 9/10 colonies originating from BAC n.3 and 10/10 colonies originating from BAC n.6 showed the correct amplification band.

For the subsequent ET cloning step, from 6 independent colonies originating from BAC n.3 (C. n.1,2,3,4,5 and 6), and 4 independent colonies originating from BAC n.6 (C. n.11, 13, 14 and 15), which still contained the recombinogenic plasmid R6K-αβγ, electrocompetent cells were prepared and electroporated with the PCR product BACGent. Positive colonies were

screened by PCR using primers GentScreenF and GentScreenR, which had been shown to yield a more robust PCR amplification than primers GentF and BACRecR.

Primer GentScreenF has the following 5'-3' sequence:

GCCTGATGCGGTATTTTCTCCT.

Primer GentScreenR has the following 5'-3' sequence:

CCTCTGTCGTTTCCTTTCTCTG.

The resulting amplification product is 1066 bp long. As with primers GentF and BACRecR, also these primers yield an amplification product only if the gentamycin cassette has been integrated in the correct site in the BAC backbone. On average, 57% colonies showed the correct band, and were therefore further characterised.

Colony n.26 (IK3H-CreEBD*-Gent-c.6/26) originating from BAC n. 3 resulted correct upon sequence analysis, and was selected as such for injection into mouse oocytes. Also in this case, as with colony O originating from BAC n.6, the targeted BAC still harboured the β-lactamase gene downstream of the Cre cassette.

X.2.5 TnpI is a novel recombinase for BAC engineering

TnpI recombinase is encoded by the Tn4430 transposon from *Bacillus thuringensis* (Mahillon and Lereclus, 1988). This is a 4149 bp transposon which codes for two proteins. A 113 K-Da protein (TnpA), homologous to the transposases of Tn3, Tn21 and Tn501, catalyses the formation of cointegrants between the donor and target replicon. The second protein (TnpI), 32 KDa, catalyses the resolution of the co-integrate and has homology to the site specific recombinases of the integrase family, like the Int recombinase of bacteriophage lambda, Cre from bacteriophage P1 and TnpA and TnpB from the Tn554 transposon.

Previous studies had established the usefulness of TnpI recombination to remove antibiotic resistance markers from genetically modified *Bacillus thuringensis* strains (Sanchis et al., 1997).

Within the internal resolution site (IRS), a 32 bp sequence was identified consisting of two inverted repeats of 14 bp each separated by a symmetric spacer of 4 nucleotides. This configuration led to the hypothesis that this site could represent the minimal recognition element for TnpI, in analogy to the loxP and FRT sites for Cre and Flp respectively. Experiments using reporter plasmids carrying a selectable marker flanked by these minimal recognition sites (called TRTs for TnpI recognition targets) demonstrated that TnpI can successfully mediate recombination between these shortened recognition elements (Youhua Deng et al., unpublished observations). The small size is of great practical relevance, because it allows to include these sites directly into oligonucleotides for any cloning purpose.

These experiments prompted to test whether the same reaction could be catalysed efficiently also on BACs, adding a novel component to the BAC engineering "toolkit". The following experiments describe the successful application of TnpI to the removal of the selectable marker cassette from the *Ikaros*-Cre modified BACs.

To this end, the same colony IK3H-CreEBD*-Gent-c.6/26 was transformed with a plasmid expressing TnpI recombinase under an arabinose inducible promoter. Cells were grown on double selection (chloramphenicol 12,5 μg/ml and kanamycin 50 μg/ml). Colonies were then picked and grown overnight in selective medium (chloramphenicol 12,5 μg/ml and kanamycin 50 μg/ml) plus L(+)arabinose (at a final concentration of 0.2%).

In principle, a fast method to detect cells which had undergone TnpI mediated recombination would have been to isolate single colonies after arabinose induction, and to replica plate them on plates with and without ampicillin respectively, since TnpI recombination would result in the loss of the β-lactamase cassette.

However, pilot experiments applying this methods systematically failed, and all colonies grew equally well on plates with and without ampicillin selection. This strongly argued that TnpI recombination was not 100% efficient, at least in BACs, and that each colony was actually a mixture of recombined and unrecombined cells. Under such circumstances, a screening method based on ampicillin sensitivity was obviously not sensitive enough, especially since the β-lactamase protein is secreted outside the cell, thereby masking the presence of cells having undergone removal of the β-lactamase cassette.

In order to overcome this problem, a PCR assay was developed based on the same primer pair described above (*Ikaros*2F and IKProbe1R). After replacement of *Ikaros* exon1 by the CreLBD* cassette, amplification with these two primers results in a product 2193 bp, which spans the TRT flanked β-lactamase cassette. After TnpI mediated recombination between the two TRT sites, amplification with these primers is expected to yield a shorter band of 1052 bp. If the mixed colony hypothesis originating from the ampicillin sensitivity experiments was correct, one would expect that colony PCR with this primer pair results in

the amplification of both unrecombined and recombined bands. This turned out indeed to be the case.

PCR was performed on liquid culture from 10 colonies grown in arabinose 0.2%. All colonies showed the expected double band (figure 49). Colonies n.1 and n.8 were then streaked to isolate colonies harbouring only the recombined BAC. From each plate, 8 colonies were picked and subjected to a second round of growth in arabinose 0.2%. PCR on these colonies showed that about half of them had lost the upper band, indicating complete TnpI mediated recombination. Colonies 8G and 1A were chosen for further characterisation. Sequence analysis was performed and confirmed that the β-lactamase gene had been precisely deleted leaving only one TRT site behind.

These experiments establish the usefulness of TnpI as a novel recombinase in BAC engineering.

Before injecting the final Tnp-deleted targeted BAC IK3H-CreEBD*-Gent-c.6/26-Tnp1A into mouse oocytes, it was necessary to obtain a pure preparation of the BAC without any contamination with the high copy plasmid expressing TnpI. In fact, in a mixed DNA preparation, the relative ratio of this high copy plasmid to the desired BAC may have dramatically decreased the chances of obtaining BAC transgenic founders. Furthermore, cointegration in the genome of the TnpI expressing cassette could have resulted in unprecedented effects.

Thus, since the TnpI expressing plasmid was not spontaneously lost from cells harbouring the targeted BAC IK3H-CreEBD*-Gent-c.6/26-Tnp1A, even after extensive culture in the absence of kanamycin selection, cultures were streaked on a double selection plate (chloramphenicol 12,5 μg/ml ampicillin 50 μg/ml). 85 colonies were picked and replica plated on plates with and without kanamycin respectively, to retrieve kanamycin sensitive

colonies. Five colonies were sensitive to kanamycin, and were therefore checked by gel electrophoresis to confirm the absence of the TnpI plasmid. All of them showed the presence of the targeted BAC without any additional plasmids. Importantly, all of them showed the same restriction pattern indicating that no undesired intramolecular recombination had occurred either during TnpI mediated recombination or during the subsequent rounds of culture to eliminate the TnpI plasmid. Thus, in the conditions tested, the targeted BAC IK3H-CreEBD*-Gent-c.6/26-Tnp1A was a very stable molecule. Colony n. 2 and n. 55 were then selected for injection into mouse oocytes.

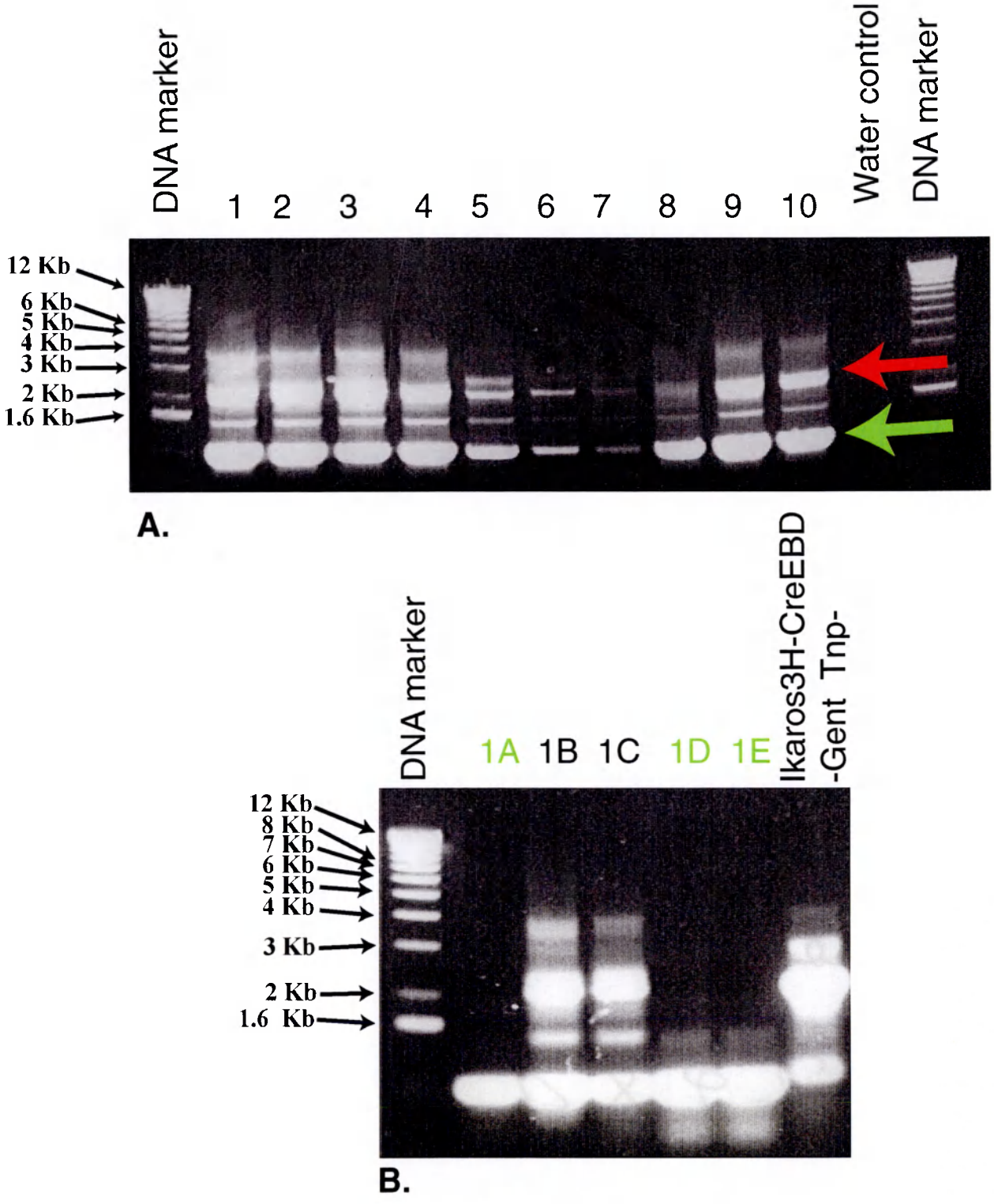


Figure 49 TnpI effectively deletes TRT flanked cassettes from BACs
A PCR assay was used to identify colonies in which the β-lactamase cassette flanked by the TnpI recognition sites (TRTs) had been deleted from the Ikaros BACs.
A. After one round of TnpI expression, all colonies show both the unrecombined (red arrow) and the recombined band (green arrow).
B. After the second round of TnpI expression, 50% of the colonies have undergone complete recombination.

X.3 Establishment of *Ikaros* Cre transgenic founder lines

Prior to oocyte microinjection, BAC DNA was prepared as described (Materials and Methods) and then quantitated by both UV spectrophotometry and gel electrophoresis. The latter was very useful to fine tune the low concentrations needed for microinjection by comparing serial dilutions with known reference standards, as described by Hammes and colleagues (Hammes et al., 2000). Following injection of the *Ikaros*-Cre BAC constructs into fertilized mouse oocytes, the following results were obtained (figures 50-52).

For the *Ikaros* BAC n.6 carrying the Cre FRTGBD*FRT cassette, 2 founders were obtained out of 9 animals born, which amounts to a frequency of 22% (figure 50). The concentration of DNA used was 2 ng/μl.

For the *Ikaros* BAC n.3 carrying the Cre FRTEBD*FRT cassette, 2 out of 3 transfers yielded founder animals. From one transfer 1 founder was identified out of 11 animals (9% frequency) (figure 51), while in the second experiment one founder was obtained out of 5 animals (20% frequency). In both cases, the concentration of DNA was 0.2ng/μl.

For the *Ikaros* BAC n.3 carrying the Cre FRTEBD*FRT cassette after TnpI recombination (colony 2), two transfers resulted in the following results: 1 founder out of 6 animals born, amounting to 16% frequency, and 2 founders out of 21 animals born, which amounts to a frequency of 9.5%. The concentration of DNA used was 2 ng/μl.

For the *Ikaros* BAC n.3 carrying the Cre FRTEBD*FRT cassette after TnpI recombination (colony 55), two transfers resulted in the following results: 1 founder out of 5 animals born, amounting to 20% frequency, and no founders out of 27 animals born. The concentration of DNA used were respectively 1 and 1.5 ng/μl.

Candidate founders were screened both by PCR and Southern blot hybridisation. PCR utilized the primers iCreF and iCreR which specifically detect the presence of the improved version of Cre recombinase. Primer iCreF has the following 5'-3' sequence:

GCCTGCCCTCCCTGTGGATGCCACCT

Primer iCreR has the following 5'-3' sequence:

GTGGCAGAAGGGGCAGCCACACCATT

Southern hybridisation was used both to confirm the PCR results and to estimate the number of copies integrated into the genome. This is an important consideration for all transgenic experiments, especially when the expression of the gene under study needs to be maintained at physiological or para-physiological levels. In this case, fidelity in expression levels was not a particular concern, since the aim was to efficiently express Cre recombinase under the *Ikaros* promoter regardless of the number of copies integrated per cell. However, it has been demonstrated that multiple integrations, which virtually always occur as tandem integrants, result in decreased expression of the transgene (Garrick et al., 1998). This is independent of the site of integration, and correlates with increased chromatin compaction and methylation at the transgene locus.

To determine the copy number, a Southern strategy was devised in which the probe detects bands of different sizes for the wild type copy of the gene and for the transgene. The probe was synthethised with primers IkProbe1F and IkProbe1R.

Primer IkProbe1F has the following 5'-3' sequence:

TCCCAAGTGAGATTAGGTAGTGTGAA

Primer IkProbe1R has the following 5'-3' sequence:

AAAATGCAAACTTACAAACAAATCAA

Following EcoRI digestion of genomic DNA from mouse tails, the probe detects a band of 1480 bp for the wild type locus, a band of 5266 for the transgene still harbouring the β-lactamase gene, and a band of 4125 for the TnpI treated transgene. As shown in figure 52 comparison of the wild type and the recombined band reveals that the various founders differ over a wide range of transgene copy number. Thus, both n. 540161 and 540178 are high copy number founders, followed by 540204, 540196, 540206, 540188, 540243 and finally 540233. 540188, 540206 and 540233 could be very likely single copy integrants.

In view of the above considerations, it will be very interesting to compare the expression levels obtained in these different lines.

These results also demonstrate that supercolied *Ikaros*-Cre BAC DNA was consistently integrated into the mouse genome, and the efficiencies were comparable to those normally reported for supercoiled plasmid DNA.

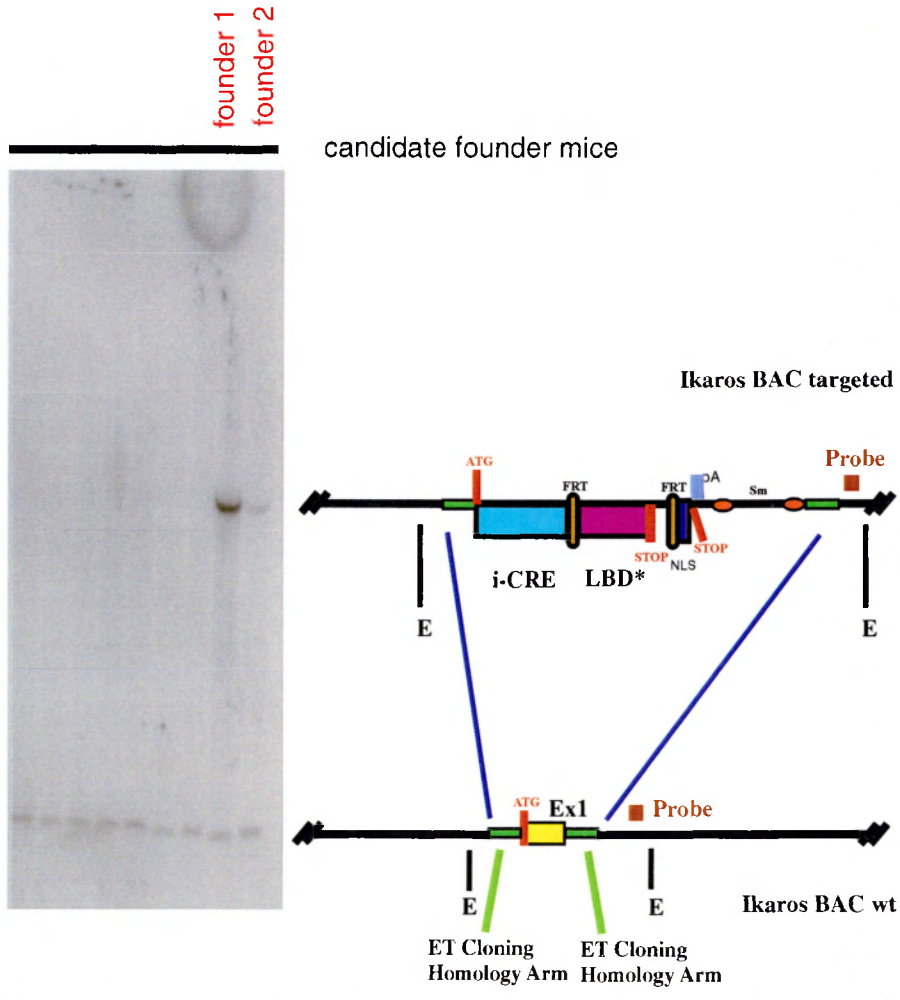


Figure 50 Southern blot screening of Ikaros-CreGBD* founders
Southern blot analysis to identify founder mice carrying the Ikaros BAC n.6 engineered with the CreFRTGBD*FRT cassette. The probe used permits to detect both the wild type and the transgenic band. From one transfer which yielded 9 anumals, 2 founder mice were obtained, amounting to a frequency of 22%.
Abbreviations: E (EcoRI); sm (selectable marker); pA (polyadenylation signal); NLS (nuclear localisation signal).

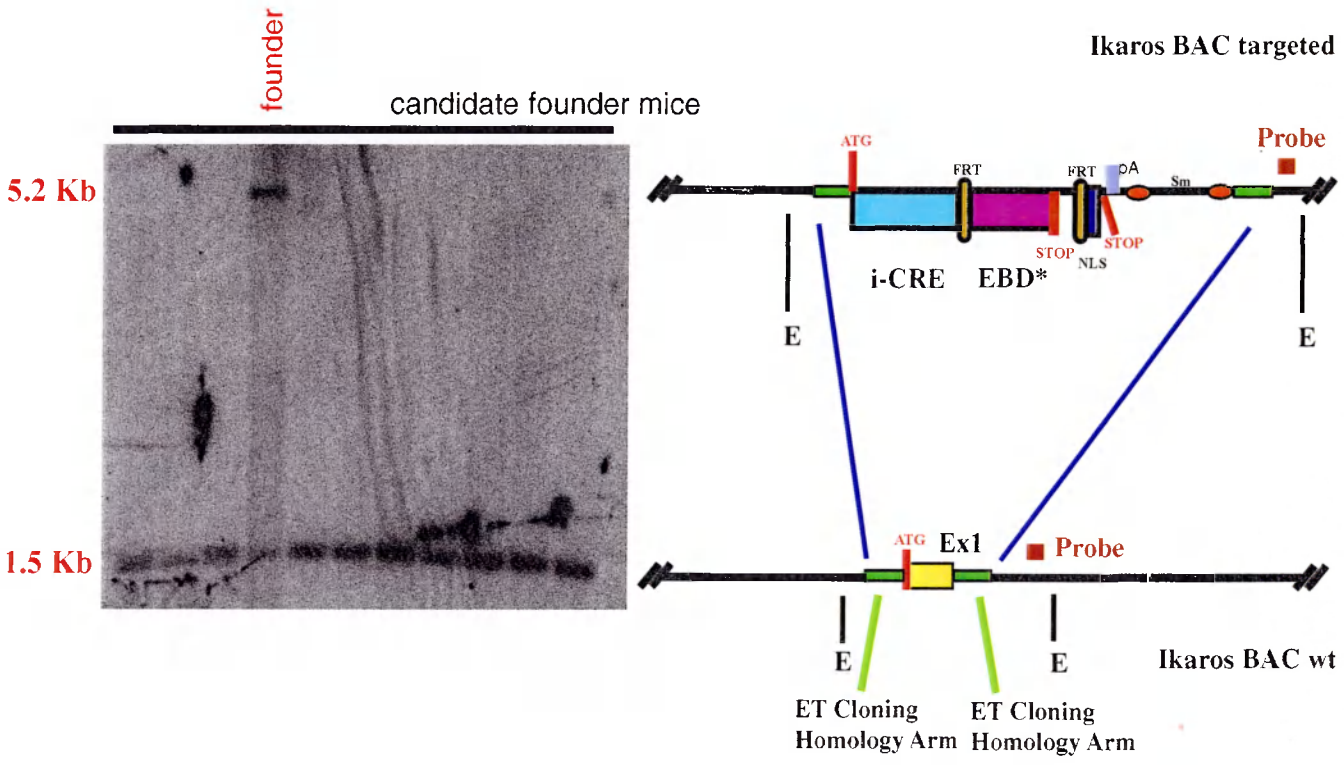


Figure 51 Southern blot screening of Ikaros-CreEBD* founders (1)
Southern blot analysis to identify founder mice carrying the Ikaros BAC n.3 engineered with the CreFRTEBD*FRT cassette. The probe used permits to detect both the wild type and the transgenic band. 1 mouse shows the recombined band, amounting to a founder frequency of 9%.
Abbreviations: E (EcoRI); sm (selectable marker); pA (polyadenylation signal); NLS (nuclear localisation signal).

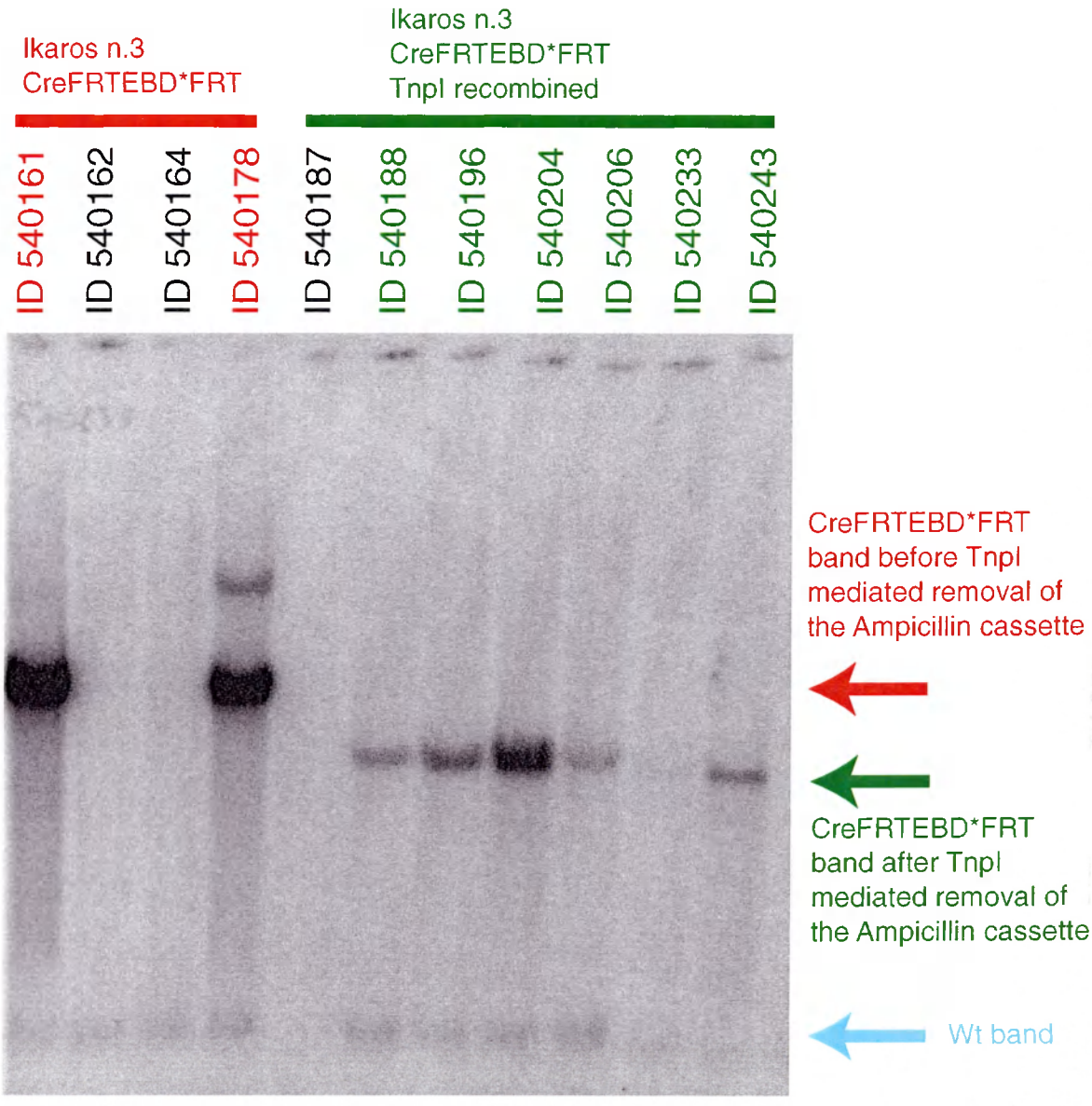


Figure 52 Southern blot screening of Ikaros-CreEBD* founders (2)
Southern blot analysis of additional founder mice carrying the Ikaros BAC n.3 engineered with the CreFRTEBD*FRT cassette.
Founders 540188-540243 have integrated the Ikaros-CreEBD* transgene in which the β-lactamase gene had been deleted by TnpI recombination, resulting in a lower band (green arrow).
The probe used permits to estimate the copy number of transgenes integrated.
ID 540161 and 540178 have integrated a very high number of copies. ID 540188, 540206 and 540233 could be single copy integrants.
See text for further details.

XI

Conclusions

The work presented in this thesis describes novel genome engineering approaches to analyse gene function in the mouse. Three mouse lines were designed with the long term objective of a completely conditional mouse model of the human leukemia associated with the t(4;11)(q21;q23) translocation.

In all three cases, the targeting constructs were assembled starting from BACs using an array of large scale cloning technologies centered on the ET recombination system.

For the *AF4* gene, a multifunctional allele was established, which can be used in the translocation model, as well as to analyse a potential loss of function phenotype and report the endogenous expression of the gene. The main results obtained from this strategy can be summarized as follows:

1) ET recombination can be used to direct simple, efficient subcloning from BACs into higher copy vectors, for a variety of downstream applications. The subcloning of the 28 kb intron 3 of the mouse *Af4* BAC was the first application of gap repair using ET recombination. This success promoted a variety of equally successful subcloning exercises across a wide range of fragment sizes from different types of target molecules.

One of these exercises was the subcloning of the desired genomic backbone for the ES cell *Af4* targeting construct, chosen after design of a Southern blot strategy to screen for homologous recombinants in ES cells. This constitutes the first example of a new way of generating mouse targeting constructs, in which the length and the location of the homology arms were chosen to suit requirements. The availability of saturating mouse BAC libraries and the expected completion of the mouse genome sequencing project will make this strategy a prime choice for engineering mouse targeting constructs. It represents a substantial

improvement over traditional approaches where elaborate cloning steps have been required.

2) A universal promoter-trap intronic knock-in cassette was developed, which can be targeted to any mouse gene regardless of its expression in ES cells. This was achieved by combining a promoterless βgeok fusion, including a splice acceptor and an IRES element, with an autonomously expressed hygromycin resistance module for genes which are not expressed in ES cells. The advantage of this approach is that as soon as the gene is activated during development or adult life, the βgeok marker reports its expression through Beta-galactosidase staining, while at the same time ablating its function by subverting its normal RNA processing. Targeting of the mouse *Af4* gene demonstrated that this double selection cassette was functional in ES cells, and that insertion of the βgeok with the SV40 late polyadenylation signal in an intron results in the absence of the wild type, full length transcript.

3) Mice homozygous for the *Af4*-LacZ allele, which results in the absence of the full length transcript, are born at the expected Mendelian frequency, indicating that, in agreement with previous reports, this gene is not essential for embryonic development. Early results with these mice, however, indicate phenotypic differences that may accord with an intron 3 truncation (this study) or an exon 11 deletion (Isnard et al., 2000).

For the *MLL* gene, a multivalent targeting construct was engineered, which provides several entry points for the functional characterisation of this gene. The main features are a protein purification tag at the aminoterminus combined with a conditional, splice-trapping cassette in intron 11. Since these two targeting sites are about 60 kb apart, two selectable markers were used. Both selectable marker cassettes were flanked by target sites for Cre and Flp. Hence, the combinatorial use of these two recombinases, in ES cells and in the mouse,

can yield four different allele configurations as a versatile platform for the following experiments: ablation of *MLL* gene function (prior to Cre and Flp recombination); isolation of the native protein complex assembled around the aminoterminal domains of MLL (after Flp recombination); isolation of the endogenous MLL protein complex, in different tissues and at different developmental stages (after Cre and Flp recombination); possible isolation of the protein complex of the leukemogenic MLL-AF4 fusion protein (after Cre and Flp recombination followed by interchromosomal translocation with the *Af4*-LoxP mouse line).

This strategy to engineer a mouse has several implications, including;

1) Due to the size of most eukaryotic genes, BACs constitute a good starting platform for the assembly of complex targeting constructs. Using ET recombination BACs can be simply engineered in two ways. First, deletion of superfluous regions from a BAC to generate a suitable backbone for the ES targeting procedure, since most unmodified BACs are likely to have homology arms of excessive length which would prevent any Southern strategy from distinguishing homologous recombinants from random integrants. Second, appropriate marker cassettes are targeted to the regions of interest within this modified BAC backbone. The results obtained for *Mll* show that 5 different selectable markers could be introduced on the same BAC, and the host bacteria could be successfully grown under simultaneous selection with all five drugs.

2) The goal of targeting via homologous recombination in ES cells two different regions of a gene of interest presents a substantial challenge. Classically, different targeting constructs have to be assembled, and the cells subjected to repeated rounds of targeting. The requirement for the two mutations to be integrated on the same allele further diminishes the fluency of this approach.

The strategy outlined here for *Mll* used a very large targeting construct (at least 65 Kb) for homologous recombination in ES cells to simultaneously target two different regions

of a gene. Two lines of evidence, the percentage of co-resistance in the double selection scheme and Southern blot analysis, indicate that the construct integrated mostly intact, without undergoing substantial breakdown; thus, these types of construct constitute a preferred route to efficiently place two mutations on the same allele. The overall frequency of homologous recombination (6.5%) is well within the range described in the literature, whose variability likely reflects the different efficiency of homologous recombination at different genomic loci.

These experiments provide both a conceptual and practical framework for the use of BACs in complex, multifunctional mouse alleles. Most modifications take place in *E. coli* and rely on the ET homologous recombination system. As shown by many experiments presented in this thesis, as well as previously reported, this methodology assures precision, fidelity and ease of manipulation in all applications of BAC engineering. Combined with the flexibility of site specific recombination, an array of mutations can be simultaneously introduced in any mouse gene, and their isolated or combined effect assessed *in vivo* in a completely conditional manner. Furthermore, a BAC vector configuration, which features two selectable marker cassettes at the two ends of the construct, represents a superior way to introduce single base substitutions in any mouse locus than previous strategies. They have been inefficient since the desired mutation, which usually had to be targeted to an exon, could not be strictly coupled to the selection marker in order not to interfere with the reading frame; under such circumstances, the frequency at which the single base mutation was correctly integrated decreased as a function of its distance from the selectable marker cassette. Recently, ET recombination was adopted to introduce single base substitutions in BACs (Muyrers et al., 2000a). In combination with the strategy described above, it becomes possible, by selecting for the integration of both ends of the construct, to ensure that a single

base substitution previously engineered into the BAC backbone is reliably targeted to the endogenous locus.

Mouse lines were established expressing Cre recombinase under the control of the *Ikaros* gene. A BAC transgenesis approach was chosen, and ET recombination was used to engineer two different *Ikaros* BACs. The following results can be outlined:

1) A novel Cre expression cassette was designed which consists of a fusion of Cre with different FRT-flanked LBD*s of nuclear hormone receptors. By appropriately regulating the expression of Flpe recombinase, this can be flexibly employed to achieve either constitutive or regulated Cre activity starting from the same original mouse line.

2) ET recombination was successfully used to target these Cre cassettes to the first exon of the *Ikaros* gene present in the BAC, as well as to delete the loxP site contained in the BAC vector. Removal of this loxP site is necessary for conditional transgenic approaches in which the random integration of a loxP in the genome is clearly undesired, and the ET based method described in this thesis is a fast and efficient way to avoid the problem.

3) TnpI recombinase from transposon Tn4430 was employed to remove the selectable marker cassette from the targeted BACs prior to oocyte injection. These experiments examined its usefulness as an additional tool for BAC engineering.

4) Following BAC injections, 10 founders were obtained, covering a wide range from single to multiple copy integrants. This should allow the establishment of lines with distinct levels of Cre expression.

XII

Materials and Methods

XII.1 Materials

XII.1.1 Enzymes

Restriction enzymes were purchased from New England Biolabs (NEB), USA or from Boehringer Mannheim, Germany. T4 DNA ligase was from NEB. Shrimp alkaline phosphatease (SAP) was from Unites States Biochemical (USB). T3 and T7 RNA polymerases, RQ1 DNAseI and recombinant human RNAsin were purchased from Promega. RNAse A, Proteinase K and BOvine Serum Albumin (BSA) were from Sigma. Amplitaq DNA polymerase and the High Fidelity PCR system were from Roche.

XII.1.2 Synthetic oligonucleotides

Synthethic oligonucleotides were synthethised by the Oligo facility of EMBL, the ARK and the Biosprings companies. When relevant, they were 5' phosphorylated or HPLC purified.

For complex ET cloning exercises, when the synthesis of particularly long oligos was required, I obtained the best results with the HPLC purified oligos from Biosprings.

XII.1.3 High Density Mouse BAC membranes

The high density mouse BAC membranes were purchased from Research Genetics, Inc, along with the reference hybridisation probe.

XII.1.4 BAC *E. coli* hosts

E.coli cells of the strain HS996 containing the BACs of interest were obtained as agar stabs from Research Genetics, Inc.

XII.1.5 Reagents for bacterial cultures

Bacterial media for routine use were prepared by the EMBL bacterial service. Antibiotics were purchased from Boehringer Mannheim. L(+)arabinose was purchased from Sigma.

XII.1.6 Mouse embryonic stem cells (ES) and Mouse embryonic fibroblasts (MEFs)

Mouse ES cells from two different lines were employed: the E14 line established in Klaus Rajewsky's lab, and the E1 line established in the lab of Austin G. Smith. Wild-type, G418 and hygromycin resistant MEFs were purchased from Genome Systems.

XII.1.7 Cell culture reagents

Reagents for the culture of mouse ES cells were purchased from Gibco-BRL (LIF, Glutamine, Sodium Pyruvate, Fetal Bovine Serum, Trypsin-EDTA), Sigma (Glucose, βMercapto-ethanol), NEN (Non essential aminoacids). Plasticware for cell culture was purchased from Falcon, Nunc, Greiner or Waki.

XII.1.8 Radioactive Isotopes

^{32}P-α-dCTP (3000Ci/mmol) and ^{32}P-α-dUTP (800Ci/mmol) were obtained from Amersham, UK.

XII.1.9 ET recombination plasmids

The ET recombination plasmids pBADabg and pR6K116/BAD/abg were described in (Muyrers et al., 1999; Zhang et al., 2000). Additional information is available at the following web address: http://www.embl-heidelberg.de/ExternalInfo/stewart/plasmids.html.

XII.2 Methods

XII.2.1 Restriction enzyme digestions

Restriction enzyme digestions of double stranded DNA were carried out according to the instructions of the suppliers, using the buffers provided. Incubation was at least 3 hours and the reaction volume was usually 50 µl.

XII.2.2 Ligations

Ligations were set up by mixing vector and insert DNA (usually with insert in at least 3 fold molar excess to vector) in a volume of 15 µl with 1 unit of DNA T4 ligase and the buffer provided by the supplier. The reaction was incubated at 16 °C for at least 12 hours.

XII.2.3 Mini preparations of plasmid DNA

Small scale DNA preparations of regular plasmids for analytical purposes were made from 1.5 ml of overnight culture and processed with the miniprep kit available from Qiagen, Germany.

For the minipreparation of BAC DNA, the protocol suggested by Research Genetics was used. Cells from a 1.5 ml overnight culture were pelleted by 30 seconds centrifugation in an eppendorf centrifuge at full speed. After medium removal, cells were resuspenden in 100 µl solution I (50mM glucose, 20 mM Tris-Cl, pH 8.0, 10 mM EDTA, pH 8.0) and mixed by pipetting. 200 µl freshly prepared solution II (0.2N NaOH, 1% SDS) were added, contents mixed and the tubes placed on ice. 150 µl solution III (KAc, pH 4.8) were added and after mixing, the cell lysates were cleared by 6' centrifugation at full speed. BAC DNA was then ethanol precipitated and the pellet washed with 70% ethanol. After complete drying, the

pellet was resuspended in 20 µl TE. Routinely, 5 µl were used for analytical restriction digestion.

XII.2.4 Maxi preparations of plasmid DNA

Large scale DNA preparations from regular plasmids were made from 500 ml overnight culture using the cartridge Maxiprep Kit available from Qiagen. Large scale preparations of BAC DNA for sequencing, analytical gel electrophoresis, oocyte injection and ES cell transfection were prepared from 500 ml overnight culture using the Large Construct Maxiprep Kit available from Qiagen, which includes an additional step of exonuclease digestion to remove all contaminant genomic DNA as well as nicked BAC molecules.

At the end of the preparation, DNA was resuspended in TE and quantitized by measuring UV absorbance at 260 nm. In the case of oocyte microinjection, in order to accurately determine very low concentrations of DNA (down to 2 ng/µl), serials dilutions of the maxiprep DNA were compared by agarose gel electrophoresis with known reference standards.

XII.2.5 Polymerase Chain Reaction

PCR reactions were carried out in a total volume of 50 µl with 0.2 mM dNTPs, primers (0.5-1 µM), amplification template (in various amounts depending on the source), Taq polymerase (2.5-5 Units/reaction) and the appropriate buffer provided by the supplier. Reaction conditions varied among different experiments, but usually conformed to the following cycling pattern: 1 minute initial denaturation at 94 °C, 30 to 35 cycles consisting of 1 minute denaturation at 94 °C, 1 minute of annealing at the relevant temperatures

(depending on the primers used) and variable times of extension at 72 °C. The last cycle was followed by 10 minutes of additional extension at 72 °C.

PCR products were purified with the PCR extraction kit from Qiagen, according to the instructions of the manufacturer.

XII.2.6 RNA extraction

RNA was extracted from mouse organs with RNAzol (Molecular Research Center, Inc.) according to the instructions of the manufacturer.

XII.2.7 Reverse transcription

Total RNA from mouse organs was used to synthethize cDNA with the Reverse Transcriptase SuperScriptII (from Gibco, BRL). In a total volume of 12 µl, 4 µl of total RNA were combined with 1 µl Oligo(dT)(500 µg/ml), and 1 µl of a dNTP mix at a concentration of 10 mM each. The mixture was heated at 65 °C for 5 minutes and quickly chilled on ice before adding the following components: 4 µl of 5X First-Strand Buffer (provided by the supplier), 2 µl of DTT (0.1 M) and 1 µl of RNAse inhibitor RNAseOUT (40 units/µl). The components were mixed and incubated at 42 °C, before adding 1 µl of Reverse Trancriptase (200 units). The reaction was allowed to proceed for 50 minutes at 42 °C. The reaction was then inactivated by heating at 70 °C for 15 minutes. To remove the RNA from the DNA-RNA hybrid, 1 µl (2 units) of *E. coli* RNAseH was added, and the reaction incubated at 37 °C for 20 minutes.

XII.2.8 Preparation and transformation of competent cells for ET cloning experiments

A detailed protocol for the preparation of competent cells for ET cloning experiments can be found at the following web address:

http://www.embl-heidelberg.de/ExternalInfo/stewart/ETprotocols.html. Here the most relevant steps are summarised.

E.Coli cells harbouring the relevant BAC were transformed with the ET expression plasmid by standard procedures and plated. Single colonies were picked and grown in 5 ml L medium overnight. 0.7ml of culture were then transferred into 70 ml of LB medium (without glucose) and grown at 37 °C shaking. Meanwhile, 10% glycerol with dH_2O was prepared, and cooled down on ice for at least 3 hours before using. When the cells reached OD600 = 0.1-0,15, 0.7ml of a 10% L-arabinose solution were added to induce ET protein expression. After a further 45-60 minutes, the cells should be at OD600 of 0.3-0.4 and ready for harvest. For BAC experiments, lower ODs (around 2.5) have actually yielded even better results. The centrifuge and the rotors SA600 or SS34 (Sorvall) were cooled by centrifuging for 10 min at -5 °C at 4,000 rpm. 35 mls of cells were spun in a first round of centrifugation for 10 min at 7,000 rpm at -5 °C, while keeping the other 35 mls on ice. After pouring away the supernatant, second 35 mls were added and respun. After pouring away supernatant, tubes were put on ice, and the cells resuspended in 5 to 10 mls ice cold 10% glycerol with an ice cold 5/10 mls pipette. 25 more mls of 10% glycerol were added and cells centrifuged. This step was repeated twice. At the end, after pouring away supernatant, the tubes were immediately dried with Kleenex tissue taking care not to touch the pellet. Cells were then resuspended in the remaining liquid (the final resuspended volume usually amounted to 100µl). 50 µl of cells were transferred into each pre-cooled eppendorf tube and frozen in

liquid N2 or used immediately. The same protocol applies, omitting the L(+) arabinose induction, for the strains like JC8679 which consitutively express the recombinogenic proteins.

For transformation, competent cells were thawed on ice and mixed with 1 to 2 µl of the recombinogenic linear DNA fragment (PCR product or a fragment excised from a plasmid). Electroporation cuvettes were precooled on ice for least 5 min. Cells were electroporated at 2.3 kV (Bio-Rad Gene Pulser) (25 µF with pulse controller set to 200 ohms), with time constants between 4.5 and 4.8. 1 ml of LB medium was added and transferrd back into the eppendorf tube. Cells were then incubated at 37 °C for 1 to 1.5 hours with shaking, and plated on suitable antibiotic selection plates.

XII.2.9 Preparation of competent cells for routine cloning

DK1, XL-1B or HB101 cells were made competent for transformation as follows. 2 mls of an overnight culture were diluted in 100 ml of L-Broth without antibiotics and grown at 37 °C up to an OD_{600} of 0.35-0.5. All subsequent steps were done at 4 °C in the cold room. Cells were spun down at 3000 rpm for 10 minutes and resuspended in 20 mls of freshly prepared, ice cold RbCl buffer (60 mM $CaCl_2$, 40 mM KAc, 15% Sucrose, 1.2 % RbCl, 0.045M $MnCl_2$). They were then spun again for 10 minutes at 300 rpm. Finally, they were resuspended in 10 ml of ice cold RbCl buffer, and 400 µl of DMSO were added. Aliquots of 200 µl were snap frozen in liquid nitrogen, and stored at -80 °C.

XII.2.10 Agarose gel electrophoresis

Agarose gel electrophoresis, used for analysis (and purification) of DNA fragments was carried out as described (Sambrook and Maniatis, 1989). The required weight of agarose for

0.4% (used for BAC restriction analysis) to 2% (w/v) gels was dissolved in 1X TAE (90 nM Tris-Cl pH 8.3, 90 mM acetic acid, 1 mM EDTA) by heating in a microwave oven. The solution was cooled and ethidium bromide was added to a concentration of 0.05 μg/ml before pouring. Alternatively, gels were stained with Ethidium Bromide (at the same concentration) after running.

Gels were run submerged in 1X TAE at variable speeds (10-15 volts/cm for minigels and 2.5-5 volts/cm for 20cmX20cm gels).

XII.2.11 Pulsed-field-gel electrophoresis

Pulsed-field-gel electrophoresis was performed on a Biorad CHEF-DRII with the following parameters: temperature (14 °C); pulse 1 (0.5 seconds); pulse 2 (10 seconds). Samples were run on a 1% agarose gel in 0.5 TBE for 18 hours at a voltage of 6 V/cm.

XII.2.12 Southern blotting

After electrophoresis, DNA gels were incubated with gentle shaking in 0.25M HCl for 10 to 30 minutes, resulting in partial depurination and facilitated transfer of the larger molecules. The gel was then transferred to 0.4N NaOH for three 15 minutes washes, to denature the DNA. The gel was then incubated in 20X SSC (3M NaCL, 0.3M NaCitrate) for 20 minutes prior to blotting onto a nylon membrane (purchased from Dupont). The gel was inverted and DNA transfer was mediated by capillary action in 20XSSC overnight. The membrane was then rinsed in 25 mM NaH_2PO_4 pH 7.2, and baked in a vacuum oven at 80 °C for at least two hours to crosslink the DNA to the membrane.

Hybridisation was performed in a glass tube in a hybridisation oven equipped with a rotating wheel. Filters were first prehybridised in Church and Gilbert buffer (7% SDS, 250 mM

NaH_2PO_4 1% BSA, 1mM EDTA) for 30 to 60 minutes incubation at the relevant temperature (65 °C to 72 °C depending on the probe), and then hybridised in the same buffer upon addition of the relevant RNA or DNA probe for 16 hours at the relevant temperature.

DNA probes were made by random prime labelling with the ready-to-go kit available from Pharmacia, according to the instructions of the manufacturer. Up tp 50 nanograms of DNA were labelled per reaction. After incorporation of radiolabelled dCTP, unincorporated DNTP were removed with purification colums (Pharmacia). Incorporation of radioactive dCTP was monitored by thin layer chromatography (TLC). For this, 0,2 µl of the reaction were spotted onto 0.1 mm cellulose polyethyleneimine TLC paper before and after incubation at 37 °C. Unincorporated and incorporated dCTP were resolved by chromatography in 0.75M KPO_4 pH 3.5, and the chromatograph dried and exposed for 5 minutes.

RNA probes were made from DNA templates carrying a T3 or T7 promoter. 1 µg of template DNA linearized with an appropriate restriction enzyme was incubated in a 15 µl volume with 3 µl transcription buffer (Promega), 2 µl 3.3 mM each ATP, GTP and CTP, 6 µl ^{32}P-α-UTP (800Ci/mmol), 1 µl RNAsin and 20 units of T3 or T7 polymerase. The reaction was incubated at 37 °C for 15 minutes, followed by addition of 40 units of RQ1 DNAse1 and further incubation at 37 °C for 15 minutes. Incorporation of UTP was monitored with TLC, as described above.

After hybridisation, the filters were rinsed with washing buffer (20 mM NaH_2PO_4 1% SDS 1mM EDTA) at room temperature, followed by two to four washes of 15 minutes each at the same temperature used for hybridisation.

Finally, autoradiography was performed by exposing the filter either to film (Kodak) or to Phosphoimager plates (Fuji).

XII.2.13 Screening of high density mouse BAC library

Hybridisation membranes from a mouse BAC library were purchased from Research Genetics, Inc. and screened according to the manufacturer's instructions. Determination of the identity of the BAC clone was performed according to the instructions of the manufacturer.

XII.2.14 BAC sequencing

Sequencing of BACs was performed at the EMBL sequencing facility (Dr. Vladimir Benes) with the direct primer walking technique described in (Benes et al., 1997).

XII.2.15 Culture of mouse ES cells and mouse embryonic fibroblasts (MEFs)

Mouse ES cells and MEFs were grown essentially as described (Matise et al., 1999). To prepare the ES cell medium, the following components were added to 500 ml of DMEM, High glucose: 90 ml of Fetal calf Serum; 6 ml of Penicillin/Streptomycin (100 units/ml), 6 ml of non essential aminoacids (10 mM), 6 ml of Sodium Pyruvate (100 mM), 6 ml of β-Mercaptoethanol (10 mM), 6 ml of L-Glutamine (200 mM) and 60 µl of LIF (from Gibco BRL) (10^7 units/ml).

XII.2.16 X-gal staining of ES cells

Cells were rinsed twice with PBS, before incubating for 2 minutes in fixative solution. They were then washed three times with PBS, before adding the staining solution and incubating at 37 °C from a few hours to overnight.

The fixative solution was prepared as follows: for a total volume of 50 ml, to 47,1 ml of PBS were added 2,7 ml of Formaldehyde (37%) and 200 µl of Gultaraldehyde (25%).

The staining solution was prepared as follows: for a total volume of 50 ml, 39.6 ml of PBS were combined with 250 µl of X-gal (200 mg/ml), 100 µl of $MgCl_2$ (1M), 5 ml of Potassium Ferricyanide (50 mM) and 5 ml of Potassium Ferrocyanide (50 mM).

XII.2.17 ES cell blastocyst injection

All ES cells injections were performed by the EMBL Transgenic service, according to standard procedures.

XII.2.18 Oocyte microinjection

All injections of BAC DNA into mouse fertilized oocytes were performed by the EMBL transgenic service according to standard procedures.

XII.2.19 Isolation of genomic DNA from ES cells and mouse tails

Genomic DNA was extracted from ES cells and mouse tails according to established methods (Matise, M.P. et al. 1999; Hammes, A. et al. 2000).

References

Aasland, R., Gibson, T. J., and Stewart, A. F. (1995). The PHD finger: implications for chromatin-mediated transcriptional regulation. Trends Biochem Sci *20*, 56-59.

Adler, H. T., Chinery, R., Wu, D. Y., Kussick, S. J., Payne, J. M., Fornace, A. J., and Tkachuk, D. C. (1999). Leukemic HRX fusion proteins inhibit GADD34-induced apoptosis and associate with the GADD34 and hSNF5/INI1 proteins. Mol Cell Biol *19*, 7050-7060.

Adler, H. T., Nallaseth, F. S., Walter, G., and Tkachuk, D. C. (1997). HRX leukemic fusion proteins form a heterocomplex with the leukemia-associated protein SET and protein phosphatase 2A. J Biol Chem *272*, 28407-28414.

Alkema, M. J., Bronk, M., Verhoeven, E., Otte, A., van't Veer, L. J., Berns, A., and van Lohuizen, M. (1997). Identification of Bmi1-interacting proteins as constituents of a multimeric mammalian polycomb complex. Genes Dev *11*, 226-240.

Angrand, P. O., Daigle, N., van der Hoeven, F., Scholer, H. R., and Stewart, A. F. (1999). Simplified generation of targeting constructs using ET recombination. Nucleic Acids Res *27*, e16.

Aplan, P. D., Chervinsky, D. S., Stanulla, M., and Burhans, W. C. (1996). Site-specific DNA cleavage within the MLL breakpoint cluster region induced by topoisomerase II inhibitors. Blood *87*, 2649-2658.

Avitahl, N., Winandy, S., Friedrich, C., Jones, B., Ge, Y., and Georgopoulos, K. (1999). Ikaros sets thresholds for T cell activation and regulates chromosome propagation. Immunity *10*, 333-343.

Barr, F. G. (1998). Translocations, cancer and the puzzle of specificity. Nat Genet *19*, 121-124.

Baskaran, K., Erfurth, F., Taborn, G., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., Iannaccone, P. M., and Domer, P. H. (1997). Cloning and developmental expression of the murine homolog of the acute leukemia proto-oncogene AF4. Oncogene *15*, 1967-1978.

Benes, V., Hostomsky, Z., Arnold, L., and Paces, V. (1993). M13 and pUC vectors with new unique restriction sites for cloning. Gene *130*, 151-152.

Benes, V., Kilger, C., Voss, H., Paabo, S., and Ansorge, W. (1997). Direct primer walking on P1 plasmid DNA. Biotechniques *23*, 98-100.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. J Mol Biol *203*, 971-983.

Bestor, T. H. (1992). Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. Embo J *11*, 2611-2617.

Beverloo, H. B., Le Coniat, M., Wijsman, J., Lillington, D. M., Bernard, O., de Klein, A., van Wering, E., Welborn, J., Young, B. D., Hagemeijer, A., and et al. (1995). Breakpoint heterogeneity in t(10;11) translocation in AML-M4/M5 resulting in fusion of AF10 and MLL is resolved by fluorescent in situ hybridization analysis. Cancer Res *55*, 4220-4224.

Biernaux, C., Loos, M., Sels, A., Huez, G., and Stryckmans, P. (1995). Detection of major bcr-abl gene expression at a very low level in blood cells of some healthy individuals. Blood *86*, 3118-3122.

Brocard, J., Feil, R., Chambon, P., and Metzger, D. (1998). A chimeric Cre recombinase inducible by synthetic,but not by natural ligands of the glucocorticoid receptor. Nucleic Acids Res *26*, 4086-4090.

Brock, H. W., and van Lohuizen, M. (2001). The Polycomb group - no longer an exclusive club? Curr Opin Genet Dev *11*, 175-181.

Brown, D., Kogan, S., Lagasse, E., Weissman, I., Alcalay, M., Pelicci, P. G., Atwater, S., and Bishop, J. M. (1997). A PMLRARalpha transgene initiates murine acute promyelocytic leukemia. Proc Natl Acad Sci U S A *94*, 2551-2556.

Brown, J. L., Mucci, D., Whiteley, M., Dirksen, M. L., and Kassis, J. A. (1998). The Drosophila Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1. Mol Cell *1*, 1057-1064.

Buchholz, F., Angrand, P. O., and Stewart, A. F. (1998). Improved properties of FLP recombinase evolved by cycling mutagenesis. Nat Biotechnol *16*, 657-662.

Buchholz, F., Refaeli, Y., Trumpp, A., and Bishop, J. M. (2000). Inducible chromosomal translocation of AML1 and ETO genes through Cre/loxP-mediated recombination in the mouse. EMBO Rep *1*, 133-139.

Cairns, B. R., Henry, N. L., and Kornberg, R. D. (1996). TFG/TAF30/ANC1, a component of the yeast SWI/SNF complex that is similar to the leukemogenic proteins ENL and AF-9. Mol Cell Biol *16*, 3308-3316.

Capili, A. D., Schultz, D. C., Rauscher, I. F., and Borden, K. L. (2001). Solution structure of the PHD domain from the KAP-1 corepressor: structural determinants for PHD, RING and LIM zinc-binding domains. Embo J *20*, 165-177.

Carmeliet, P., Ferreira, V., Breier, G., Pollefeyt, S., Kieckens, L., Gertsenstein, M., Fahrig, M., Vandenhoeck, A., Harpal, K., Eberhardt, C., *et al.* (1996). Abnormal blood vessel development and lethality in embryos lacking a single VEGF allele. Nature *380*, 435-439.

Caslini, C., Alarcon, A. S., Hess, J. L., Tanaka, R., Murti, K. G., and Biondi, A. (2000a). The amino terminus targets the mixed lineage leukemia (MLL) protein to the nucleolus, nuclear matrix and mitotic chromosomal scaffolds. Leukemia *14*, 1898-1908.

Caslini, C., Shilatifard, A., Yang, L., and Hess, J. L. (2000b). The amino terminus of the mixed lineage leukemia protein (MLL) promotes cell cycle arrest and monocytic differentiation. Proc Natl Acad Sci U S A *97*, 2797-2802.

Caspersson, T., Zech, L., and Johansson, C. (1970). Differential binding of alkylating fluorochromes in human chromosomes. Exp Cell Res *60*, 315-319.

Castellanos, A., Pintado, B., Weruaga, E., Arevalo, R., Lopez, A., Orfao, A., and Sanchez-Garcia, I. (1997). A BCR-ABL(p190) fusion gene made by homologous recombination causes B-cell acute lymphoblastic leukemias in chimeric mice with independence of the endogenous bcr product. Blood *90*, 2168-2174.

Castilla, L. H., Wijmenga, C., Wang, Q., Stacy, T., Speck, N. A., Eckhaus, M., Marin-Padilla, M., Collins, F. S., Wynshaw-Boris, A., and Liu, P. P. (1996). Failure of embryonic hematopoiesis and lethal hemorrhages in mouse embryos heterozygous for a knocked-in leukemia gene CBFB-MYH11. Cell *87*, 687-696.

Chakrabarti, L., Bristulf, J., Foss, G. S., and Davies, K. E. (1998). Expression of the murine homologue of FMR2 in mouse brain and during development. Hum Mol Genet *7*, 441-448.

Chaplin, T., Ayton, P., Bernard, O. A., Saha, V., Della Valle, V., Hillion, J., Gregorini, A., Lillington, D., Berger, R., and Young, B. D. (1995a). A novel class of zinc finger/leucine zipper genes identified from the molecular cloning of the t(10;11) translocation in acute leukemia. Blood *85*, 1435-1441.

Chaplin, T., Bernard, O., Beverloo, H. B., Saha, V., Hagemeijer, A., Berger, R., and Young, B. D. (1995b). The t(10;11) translocation in acute myeloid leukemia (M5)

consistently fuses the leucine zipper motif of AF10 onto the HRX gene. Blood *86*, 2073-2076.

Chaplin, T., Jones, L., Debernardi, S., Hill, A. S., Lillington, D. M., and Young, B. D. (2001). Molecular analysis of the genomic inversion and insertion of AF10 into MLL suggests a single-step event. Genes Chromosomes Cancer *30*, 175-180.

Chen, C. S., Hilden, J. M., Frestedt, J., Domer, P. H., Moore, R., Korsmeyer, S. J., and Kersey, J. H. (1993). The chromosome 4q21 gene (AF-4/FEL) is widely expressed in normal tissues and shows breakpoint diversity in t(4;11)(q21;q23) acute leukemia. Blood *82*, 1080-1085.

Cimino, G., Moir, D. T., Canaani, O., Williams, K., Crist, W. M., Katzav, S., Cannizzaro, L., Lange, B., Nowell, P. C., Croce, C. M., and et al. (1991). Cloning of ALL-1, the locus involved in leukemias with the t(4;11)(q21;q23), t(9;11)(p22;q23), and t(11;19)(q23;p13) chromosome translocations. Cancer Res *51*, 6712-6714.

Cimino, G., Nakamura, T., Gu, Y., Canaani, O., Prasad, R., Crist, W. M., Carroll, A. J., Baer, M., Bloomfield, C. D., Nowell, P. C., and et al. (1992). An altered 11-kilobase transcript in leukemic cell lines with the t(4;11)(q21;q23) chromosome translocation. Cancer Res *52*, 3811-3813.

Cimino, G., Rapanotti, M. C., Biondi, A., Elia, L., Lo Coco, F., Price, C., Rossi, V., Rivolta, A., Canaani, E., Croce, C. M., *et al.* (1997). Infant acute leukemias show the same biased distribution of ALL1 gene breaks as topoisomerase II related secondary acute leukemias. Cancer Res *57*, 2879-2883.

Collins, E. C., Pannell, R., Simpson, E. M., Forster, A., and Rabbitts, T. H. (2000). Inter-chromosomal recombination of Mll and Af9 genes mediated by cre-loxP in mouse development. EMBO Rep *1*, 127-132.

Collins, R. T., and Treisman, J. E. (2000). Osa-containing Brahma chromatin remodeling complexes are required for the repression of wingless target genes. Genes Dev *14*, 3140-3152.

Corral, J., Forster, A., Thompson, S., Lampert, F., Kaneko, Y., Slater, R., Kroes, W. G., van der Schoot, C. E., Ludwig, W. D., Karpas, A., and et al. (1993). Acute leukemias of different lineages have similar MLL gene fusions encoding related chimeric proteins resulting from chromosomal translocation. Proc Natl Acad Sci U S A *90*, 8538-8542.

Corral, J., Lavenir, I., Impey, H., Warren, A. J., Forster, A., Larson, T. A., Bell, S., McKenzie, A. N., King, G., and Rabbitts, T. H. (1996). An Mll-AF9 fusion gene made by

homologous recombination causes acute leukemia in chimeric mice: a method to create fusion oncogenes. Cell *85*, 853-861.

Cortes, M., Wong, E., Koipally, J., and Georgopoulos, K. (1999). Control of lymphocyte development by the Ikaros gene family. Curr Opin Immunol *11*, 167-171.

Cremer, T., and Cremer, C. (2001). Chromosome Territories, Nuclear Architecture and Gene Regulation in Mammalian Cells. Nat Rev Genet *2*, 292-301.

Cui, X., De Vivo, I., Slany, R., Miyamoto, A., Firestein, R., and Cleary, M. L. (1998). Association of SET domain and myotubularin-related proteins modulates growth control. Nat Genet *18*, 331-337.

Daley, G. Q., McLaughlin, J., Witte, O. N., and Baltimore, D. (1987). The CML-specific P210 bcr/abl protein, unlike v-abl, does not transform NIH/3T3 fibroblasts. Science *237*, 532-535.

Danielian, P. S., Muccino, D., Rowitch, D. H., Michael, S. K., and McMahon, A. P. (1998). Modification of gene activity in mouse embryos in utero by a tamoxifen-inducible form of Cre recombinase. Curr Biol *8*, 1323-1326.

Danielian, P. S., White, R., Hoare, S. A., Fawell, S. E., and Parker, M. G. (1993). Identification of residues in the estrogen receptor that confer differential sensitivity to estrogen and hydroxytamoxifen. Mol Endocrinol *7*, 232-240.

David, G., Terris, B., Marchio, A., Lavau, C., and Dejean, A. (1997). The acute promyelocytic leukemia PML-RAR alpha protein induces hepatic preneoplastic and neoplastic lesions in transgenic mice. Oncogene *14*, 1547-1554.

Delmas, V., Stokes, D. G., and Perry, R. P. (1993). A mammalian DNA-binding protein that contains a chromodomain and an SNF2/SWI2-like helicase domain. Proc Natl Acad Sci U S A *90*, 2414-2418.

Dimartino, J. F., and Cleary, M. L. (1999). Mll rearrangements in haematological malignancies: lessons from clinical and biological studies. Br J Haematol *106*, 614-626.

Djabali, M., Selleri, L., Parry, P., Bower, M., Young, B. D., and Evans, G. A. (1992). A trithorax-like gene is interrupted by chromosome 11q23 translocations in acute leukaemias. Nat Genet *2*, 113-118.

Dobson, C. L., Warren, A. J., Pannell, R., Forster, A., Lavenir, I., Corral, J., Smith, A. J., and Rabbitts, T. H. (1999). The mll-AF9 gene fusion in mice controls myeloproliferation and specifies acute myeloid leukaemogenesis. Embo J *18*, 3564-3574.

Dobson, C. L., Warren, A. J., Pannell, R., Forster, A., and Rabbitts, T. H. (2000). Tumorigenesis in mice with a fusion of the leukaemia oncogene Mll and the bacterial lacZ gene. Embo J *19*, 843-851.

Dolken, G., Illerhaus, G., Hirt, C., and Mertelsmann, R. (1996). BCL-2/JH rearrangements in circulating B cells of healthy blood donors and patients with nonmalignant diseases. J Clin Oncol *14*, 1333-1344.

Domer, P. H., Fakharzadeh, S. S., Chen, C. S., Jockel, J., Johansen, L., Silverman, G. A., Kersey, J. H., and Korsmeyer, S. J. (1993). Acute mixed-lineage leukemia t(4;11)(q21;q23) generates an MLL-AF4 fusion product. Proc Natl Acad Sci U S A *90*, 7884-7888.

Domer, P. H., Head, D. R., Renganathan, N., Raimondi, S. C., Yang, E., and Atlas, M. (1995). Molecular analysis of 13 cases of MLL/11q23 secondary acute leukemia and identification of topoisomerase II consensus-binding sequences near the chromosomal breakpoint of a secondary leukemia with the t(4;11). Leukemia *9*, 1305-1312.

Duncan, I. (1987). The bithorax complex. Annu Rev Genet *21*, 285-319.

Early, E., Moore, M. A., Kakizuka, A., Nason-Burchenal, K., Martin, P., Evans, R. M., and Dmitrovsky, E. (1996). Transgenic expression of PML/RARalpha impairs myelopoiesis. Proc Natl Acad Sci U S A *93*, 7900-7904.

Ernst, P., Wang, J., Huang, M., Goodman, R. H., and Korsmeyer, S. J. (2001). Mll and creb bind cooperatively to the nuclear coactivator creb-binding protein. Mol Cell Biol *21*, 2249-2258.

Farkas, G., Gausz, J., Galloni, M., Reuter, G., Gyurkovics, H., and Karch, F. (1994). The Trithorax-like gene encodes the Drosophila GAGA factor. Nature *371*, 806-808.

Fauvarque, M. O., Zuber, V., and Dura, J. M. (1995). Regulation of polyhomeotic transcription may involve local changes in chromatin activity in Drosophila. Mech Dev *52*, 343-355.

Feil, R., Brocard, J., Mascrez, B., LeMeur, M., Metzger, D., and Chambon, P. (1996). Ligand-activated site-specific recombination in mice. Proc Natl Acad Sci U S A *93*, 10887-10890.

Feil, R., Wagner, J., Metzger, D., and Chambon, P. (1997). Regulation of Cre recombinase activity by mutated estrogen receptor ligand-binding domains. Biochem Biophys Res Commun *237*, 752-757.

Ferrucci, P. F., Grignani, F., Pearson, M., Fagioli, M., Nicoletti, I., and Pelicci, P. G. (1997). Cell death induction by the acute promyelocytic leukemia-specific PML/RARalpha fusion protein. Proc Natl Acad Sci U S A *94*, 10901-10906.

Fiering, S., Kim, C. G., Epner, E. M., and Groudine, M. (1993). An "in-out" strategy using gene targeting and FLP recombinase for the functional dissection of complex DNA regulatory elements: analysis of the beta-globin locus control region. Proc Natl Acad Sci U S A *90*, 8469-8473.

FitzGerald, K. T., and Diaz, M. O. (1999). MLL2: A new mammalian member of the trx/MLL family of genes. Genomics *59*, 187-192.

Ford, A. M., Ridge, S. A., Cabrera, M. E., Mahmoud, H., Steel, C. M., Chan, L. C., and Greaves, M. (1993). In utero rearrangements in the trithorax-related oncogene in infant leukaemias. Nature *363*, 358-360.

Fu, X., and Kamps, M. P. (1997). E2a-Pbx1 induces aberrant expression of tissue-specific and developmentally regulated genes when expressed in NIH 3T3 fibroblasts. Mol Cell Biol *17*, 1503-1512.

Fuks, F., Burgers, W. A., Brehm, A., Hughes-Davies, L., and Kouzarides, T. (2000). DNA methyltransferase Dnmt1 associates with histone deacetylase activity. Nat Genet *24*, 88-91.

Gale, K. B., Ford, A. M., Repp, R., Borkhardt, A., Keller, C., Eden, O. B., and Greaves, M. F. (1997). Backtracking leukemia to birth: identification of clonotypic gene fusion sequences in neonatal blood spots. Proc Natl Acad Sci U S A *94*, 13950-13954.

Garrick, D., Fiering, S., Martin, D. I., and Whitelaw, E. (1998). Repeat-induced gene silencing in mammals. Nat Genet *18*, 56-59.

Gecz, J., Bielby, S., Sutherland, G. R., and Mulley, J. C. (1997). Gene structure and subcellular localization of FMR2, a member of a new family of putative transcription activators. Genomics *44*, 201-213.

Gecz, J., Gedeon, A. K., Sutherland, G. R., and Mulley, J. C. (1996). Identification of the gene FMR2, associated with FRAXE mental retardation. Nat Genet *13*, 105-108.

Georgopoulos, K., Bigby, M., Wang, J. H., Molnar, A., Wu, P., Winandy, S., and Sharpe, A. (1994). The Ikaros gene is required for the development of all lymphoid lineages. Cell *79*, 143-156.

Georgopoulos, K., Moore, D. D., and Derfler, B. (1992). Ikaros, an early lymphoid-specific transcription factor and a putative mediator for T cell commitment. Science *258*, 808-812.

Georgopoulos, K., Winandy, S., and Avitahl, N. (1997). The role of the Ikaros gene in lymphocyte development and homeostasis. Annu Rev Immunol *15*, 155-176.

Gildea, J. J., Lopez, R., and Shearn, A. (2000). A screen for new trithorax group genes identified little imaginal discs, the Drosophila melanogaster homologue of human retinoblastoma binding protein 2. Genetics *156*, 645-663.

Gillert, E., Leis, T., Repp, R., Reichel, M., Hosch, A., Breitenlohner, I., Angermuller, S., Borkhardt, A., Harbott, J., Lampert, F., *et al.* (1999). A DNA damage repair mechanism is involved in the origin of chromosomal translocations t(4;11) in primary leukemic cells. Oncogene *18*, 4663-4671.

Gilley, J., and Fried, M. (1999). Extensive gene order differences within regions of conserved synteny between the Fugu and human genomes: implications for chromosomal evolution and the cloning of disease genes. Hum Mol Genet *8*, 1313-1320.

Golic, K. G. (1991). Site-specific recombination between homologous chromosomes in Drosophila. Science *252*, 958-961.

Golic, M. M., and Golic, K. G. (1996). A quantitative measure of the mitotic pairing of alleles in Drosophila melanogaster and the influence of structural heterozygosity. Genetics *143*, 385-400.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531-537.

Gregorini, A., Sahin, F. I., Lillington, D. M., Meerabux, J., Saha, V., McCullagh, P., Bocci, M., Menevse, S., Papa, S., and Young, B. D. (1996). Gene BR140, which is related to AF10 and AF17, maps to chromosome band 3p25. Genes Chromosomes Cancer *17*, 269-272.

Grignani, F., Ferrucci, P. F., Testa, U., Talamo, G., Fagioli, M., Alcalay, M., Mencarelli, A., Peschle, C., Nicoletti, I., and et al. (1993). The acute promyelocytic leukemia-specific PML-RAR alpha fusion protein inhibits differentiation and promotes survival of myeloid precursor cells. Cell *74*, 423-431.

Grisolano, J. L., Wesselschmidt, R. L., Pelicci, P. G., and Ley, T. J. (1997). Altered myeloid development and acute leukemia in transgenic mice expressing PML-RAR alpha under control of cathepsin G regulatory sequences. Blood *89*, 376-387.

Gu, Y., Alder, H., Nakamura, T., Schichman, S. A., Prasad, R., Canaani, O., Saito, H., Croce, C. M., and Canaani, E. (1994). Sequence analysis of the breakpoint cluster region in the ALL-1 gene involved in acute leukemia. Cancer Res *54*, 2326-2330.

Gu, Y., Cimino, G., Alder, H., Nakamura, T., Prasad, R., Canaani, O., Moir, D. T., Jones, C., Nowell, P. C., Croce, C. M., and et al. (1992a). The (4;11)(q21;q23) chromosome translocations in acute leukemias involve the VDJ recombinase. Proc Natl Acad Sci U S A *89*, 10464-10468.

Gu, Y., Nakamura, T., Alder, H., Prasad, R., Canaani, O., Cimino, G., Croce, C. M., and Canaani, E. (1992b). The t(4;11) chromosome translocation of human acute leukemias fuses the ALL-1 gene, related to Drosophila trithorax, to the AF-4 gene. Cell *71*, 701-708.

Gu, Y., Shen, Y., Gibbs, R. A., and Nelson, D. L. (1996). Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. Nat Genet *13*, 109-113.

Hagstrom, K., Muller, M., and Schedl, P. (1997). A Polycomb and GAGA dependent silencer adjoins the Fab-7 boundary in the Drosophila bithorax complex. Genetics *146*, 1365-1380.

Hahm, K., Cobb, B. S., McCarty, A. S., Brown, K. E., Klug, C. A., Lee, R., Akashi, K., Weissman, I. L., Fisher, A. G., and Smale, S. T. (1998). Helios, a T cell-restricted Ikaros family member that quantitatively associates with Ikaros at centromeric heterochromatin. Genes Dev *12*, 782-796.

Hahm, K., Ernst, P., Lo, K., Kim, G. S., Turck, C., and Smale, S. T. (1994). The lymphoid transcription factor LyF-1 is encoded by specific, alternatively spliced mRNAs derived from the Ikaros gene. Mol Cell Biol *14*, 7111-7123.

Hamilton, C. M., Aldea, M., Washburn, B. K., Babitzke, P., and Kushner, S. R. (1989). New method for generating deletions and gene replacements in Escherichia coli. J Bacteriol *171*, 4617-4622.

Hanson, R. D., Hess, J. L., Yu, B. D., Ernst, P., van Lohuizen, M., Berns, A., van der Lugt, N. M., Shashikant, C. S., Ruddle, F. H., Seto, M., and Korsmeyer, S. J. (1999). Mammalian Trithorax and polycomb-group homologues are antagonistic regulators of homeotic development. Proc Natl Acad Sci U S A *96*, 14372-14377.

Hasty, P., Ramirez-Solis, R., Krumlauf, R., and Bradley, A. (1991). Introduction of a subtle mutation into the Hox-2.6 locus in embryonic stem cells. Nature *350*, 243-246.

He, L. Z., Guidez, F., Tribioli, C., Peruzzi, D., Ruthardt, M., Zelent, A., and Pandolfi, P. P. (1998). Distinct interactions of PML-RARalpha and PLZF-RARalpha with co-repressors determine differential responses to RA in APL. Nat Genet *18*, 126-135.

Heisterkamp, N., Jenster, G., Kioussis, D., Pattengale, P. K., and Groffen, J. (1991). Human bcr-abl gene has a lethal effect on embryogenesis. Transgenic Res *1*, 45-53.

Hendrich, B., Abbott, C., McQueen, H., Chambers, D., Cross, S., and Bird, A. (1999). Genomic structure and chromosomal mapping of the murine and human Mbd1, Mbd2, Mbd3, and Mbd4 genes. Mamm Genome *10*, 906-912.

Herault, Y., Rassoulzadegan, M., Cuzin, F., and Duboule, D. (1998). Engineering chromosomes in mice through targeted meiotic recombination (TAMERE). Nat Genet *20*, 381-384.

Hess, J. L., Yu, B. D., Li, B., Hanson, R., and Korsmeyer, S. J. (1997). Defects in yolk sac hematopoiesis in Mll-null embryos. Blood *90*, 1799-1806.

Hunger, S. P., and Cleary, M. L. (1998). What significance should we attribute to the detection of MLL fusion transcripts? Blood *92*, 709-711.

Huntsman, D. G., Chin, S. F., Muleris, M., Batley, S. J., Collins, V. P., Wiedemann, L. M., Aparicio, S., and Caldas, C. (1999). MLL2, the second human homolog of the Drosophila trithorax gene, maps to 19q13.1 and is amplified in solid tumor cell lines. Oncogene *18*, 7975-7984.

Huth, J. R., Bewley, C. A., Nissen, M. S., Evans, J. N., Reeves, R., Gronenborn, A. M., and Clore, G. M. (1997). The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. Nat Struct Biol *4*, 657-665.

Ida, K., Kitabayashi, I., Taki, T., Taniwaki, M., Noro, K., Yamamoto, M., Ohki, M., and Hayashi, Y. (1997). Adenoviral E1A-associated protein p300 is involved in acute myeloid leukemia with t(11;22)(q23;q13). Blood *90*, 4699-4704.

Indra, A. K., Warot, X., Brocard, J., Bornert, J. M., Xiao, J. H., Chambon, P., and Metzger, D. (1999). Temporally-controlled site-specific mutagenesis in the basal layer of the epidermis: comparison of the recombinase activity of the tamoxifen-inducible Cre-ER(T) and Cre-ER(T2) recombinases. Nucleic Acids Res *27*, 4324-4327.

Ingham, P. W. (1981). trithorax: A new homeotic mutation of *Drosophila melanogaster*. II. Roux Arch. Dev. Biol. 190:365-369.

Ingham, P. W. (1998). trithorax and the regulation of homeotic gene expression in Drosophila: a historical perspective. Int J Dev Biol *42*, 423-429.

Ioannou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., Batzer, M. A., and de Jong, P. J. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. Nat Genet *6*, 84-89.

Isnard, P., Core, N., Naquet, P., and Djabali, M. (2000). Altered lymphoid development in mice deficient for the mAF4 proto-oncogene. Blood *96*, 705-710.

Isnard, P., Depetris, D., Mattei, M. G., Ferrier, P., and Djabali, M. (1998). cDNA cloning, expression and chromosomal localization of the murine AF-4 gene involved in human leukemia. Mamm Genome *9*, 1065-1068.

Ito, T., Bulger, M., Kobayashi, R., and Kadonaga, J. T. (1996). Drosophila NAP-1 is a core histone chaperone that functions in ATP-facilitated assembly of regularly spaced nucleosomal arrays. Mol Cell Biol *16*, 3112-3124.

Jacobson, R. H., Zhang, X. J., DuBose, R. F., and Matthews, B. W. (1994). Three-dimensional structure of beta-galactosidase from E. coli. Nature *369*, 761-766.

Janssen, J. W., Ludwig, W. D., Borkhardt, A., Spadinger, U., Rieder, H., Fonatsch, C., Hossfeld, D. K., Harbott, J., Schulz, A. S., Repp, R., and et al. (1994). Pre-pre-B acute lymphoblastic leukemia: high frequency of alternatively spliced ALL1-AF4 transcripts and absence of minimal residual disease during complete remission. Blood *84*, 3835-3842.

Jenuwein, T., Laible, G., Dorn, R., and Reuter, G. (1998). SET domain proteins modulate chromatin domains in eu- and heterochromatin. Cell Mol Life Sci *54*, 80-93.

Jessen, J. R., Meng, A., McFarlane, R. J., Paw, B. H., Zon, L. I., Smith, G. R., and Lin, S. (1998). Modification of bacterial artificial chromosomes through chi-stimulated homologous recombination and its application in zebrafish transgenesis. Proc Natl Acad Sci U S A *95*, 5121-5126.

Johansson, B., Moorman, A. V., Haas, O. A., Watmore, A. E., Cheung, K. L., Swanton, S., and Secker-Walker, L. M. (1998). Hematologic malignancies with t(4;11)(q21;q23)--a cytogenetic, morphologic, immunophenotypic and clinical study of 183 cases. European 11q23 Workshop participants. Leukemia *12*, 779-787.

Johnson, L., Mercer, K., Greenbaum, D., Bronson, R. T., Crowley, D., Tuveson, D. A., and Jacks, T. (2001). Somatic activation of the K-ras oncogene causes early onset lung cancer in mice. Nature *410*, 1111-1116.

Jones, P. L., Veenstra, G. J., Wade, P. A., Vermaak, D., Kass, S. U., Landsberger, N., Strouboulis, J., and Wolffe, A. P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet *19*, 187-191.

Kaufman, T. C., Seeger, M. A., and Olsen, G. (1990). Molecular and genetic organization of the antennapedia gene complex of Drosophila melanogaster. Adv Genet *27*, 309-362.

Kawasaki, H., Eckner, R., Yao, T. P., Taira, K., Chiu, R., Livingston, D. M., and Yokoyama, K. K. (1998). Distinct roles of the co-activators p300 and CBP in retinoic-acid-induced F9-cell differentiation. Nature *393*, 284-289.

Kazmierczak, B., Bullerdiek, J., Pham, K. H., Bartnitzke, S., and Wiesner, H. (1998a). Intron 3 of HMGIC is the most frequent target of chromosomal aberrations in human tumors and has been conserved basically for at least 30 million years. Cancer Genet Cytogenet *103*, 175-177.

Kazmierczak, B., Dal Cin, P., Wanschura, S., Borrmann, L., Fusco, A., Van den Berghe, H., and Bullerdiek, J. (1998b). HMGIY is the target of 6p21.3 rearrangements in various benign mesenchymal tumors. Genes Chromosomes Cancer *23*, 279-285.

Kazmierczak, B., Meyer-Bolte, K., Tran, K. H., Wockel, W., Breightman, I., Rosigkeit, J., Bartnitzke, S., and Bullerdiek, J. (1999). A high frequency of tumors with rearrangements of genes of the HMGI(Y) family in a series of 191 pulmonary chondroid hamartomas. Genes Chromosomes Cancer *26*, 125-133.

Kellendonk, C., Tronche, F., Monaghan, A. P., Angrand, P. O., Stewart, F., and Schutz, G. (1996). Regulation of Cre recombinase activity by the synthetic steroid RU 486. Nucleic Acids Res *24*, 1404-1411.

Kelley, C. M., Ikeda, T., Koipally, J., Avitahl, N., Wu, L., Georgopoulos, K., and Morgan, B. A. (1998). Helios, a novel dimerization partner of Ikaros expressed in the earliest hematopoietic progenitors. Curr Biol *8*, 508-515.

Kennison, J. A. (1995). The Polycomb and trithorax group proteins of Drosophila: trans-regulators of homeotic gene function. Annu Rev Genet *29*, 289-303.

Kennison, J. A., and Tamkun, J. W. (1988). Dosage-dependent modifiers of polycomb and antennapedia mutations in Drosophila. Proc Natl Acad Sci U S A *85*, 8136-8140.

Kennison, J. A., and Tamkun, J. W. (1992). Trans-regulation of homeotic genes in Drosophila. New Biol *4*, 91-96.

Kersey, J. H., Wang, D., and Oberto, M. (1998). Resistance of t(4;11) (MLL-AF4 fusion gene) leukemias to stress-induced cell death: possible mechanism for extensive extramedullary accumulation of cells and poor prognosis. Leukemia *12*, 1561-1564.

Kim, C. G., Epner, E. M., Forrester, W. C., and Groudine, M. (1992). Inactivation of the human beta-globin gene by targeted insertion into the beta-globin locus control region. Genes Dev *6*, 928-938.

Kim, J., Sif, S., Jones, B., Jackson, A., Koipally, J., Heller, E., Winandy, S., Viel, A., Sawyer, A., Ikeda, T., *et al.* (1999). Ikaros DNA-binding proteins direct formation of chromatin remodeling complexes in lymphocytes. Immunity *10*, 345-355.

Kim-Rouille, M. H., MacGregor, A., Wiedemann, L. M., Greaves, M. F., and Navarrete, C. (1999). MLL-AF4 gene fusions in normal newborns. Blood *93*, 1107-1108.

Koipally, J., Kim, J., Jones, B., Jackson, A., Avitahl, N., Winandy, S., Trevisan, M., Nichogiannopoulou, A., Kelley, C., and Georgopoulos, K. (1999a). Ikaros chromatin remodeling complexes in the control of differentiation of the hemo-lymphoid system. Cold Spring Harb Symp Quant Biol *64*, 79-86.

Koipally, J., Renold, A., Kim, J., and Georgopoulos, K. (1999b). Repression by Ikaros and Aiolos is mediated through histone deacetylase complexes. Embo J *18*, 3090-3100.

Kolodner, R., Hall, S. D., and Luisi-DeLuca, C. (1994). Homologous pairing proteins encoded by the Escherichia coli recE and recT genes. Mol Microbiol *11*, 23-30.

Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell *44*, 283-292.

Kozak, M. (1987). At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J Mol Biol *196*, 947-950.

Kroon, E., Krosl, J., Thorsteinsdottir, U., Baban, S., Buchberg, A. M., and Sauvageau, G. (1998). Hoxa9 transforms primary bone marrow cells through specific collaboration with Meis1a but not Pbx1b. Embo J *17*, 3714-3725.

Kroon, E., Thorsteinsdottir, U., Mayotte, N., Nakamura, T., and Sauvageau, G. (2001). NUP98-HOXA9 expression in hemopoietic stem cells induces chronic and acute myeloid leukemias in mice. Embo J *20*, 350-361.

Kuroda, M., Ishida, T., Takanashi, M., Satoh, M., Machinami, R., and Watanabe, T. (1997). Oncogenic transformation and inhibition of adipocytic conversion of preadipocytes by TLS/FUS-CHOP type II chimeric protein. Am J Pathol *151*, 735-744.

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., and Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature *410*, 116-120.

LaJeunesse, D., and Shearn, A. (1996). E(z): a polycomb group gene or a trithorax group gene? Development *122*, 2189-2197.

Lallemand, Y., Luria, V., Haffner-Krausz, R., and Lonai, P. (1998). Maternally expressed PGK-Cre transgene as a tool for early and uniform activation of the Cre site-specific recombinase. Transgenic Res *7*, 105-112.

Lamond, A. I., and Earnshaw, W. C. (1998). Structure and function in the nucleus. Science *280*, 547-553.

Lavau, C., Szilvassy, S. J., Slany, R., and Cleary, M. L. (1997). Immortalization and leukemic transformation of a myelomonocytic precursor by retrovirally transduced HRX-ENL. Embo J *16*, 4226-4237.

Lawrence, H. J., Rozenfeld, S., Cruz, C., Matsukuma, K., Kwong, A., Komuves, L., Buchberg, A. M., and Largman, C. (1999). Frequent co-expression of the HOXA9 and MEIS1 homeobox genes in human myeloid leukemias. Leukemia *13*, 1993-1999.

Lawrence, H. J., Sauvageau, G., Humphries, R. K., and Largman, C. (1996). The role of HOX homeobox genes in normal and leukemic hematopoiesis. Stem Cells *14*, 281-291.

Li, M., Makkinje, A., and Damuni, Z. (1996). The myeloid leukemia-associated protein SET is a potent inhibitor of protein phosphatase 2A. J Biol Chem *271*, 11059-11062.

Li, Q., Frestedt, J. L., and Kersey, J. H. (1998). AF4 encodes a ubiquitous protein that in both native and MLL-AF4 fusion types localizes to subnuclear compartments. Blood *92*, 3841-3847.

Linder, B., Newman, R., Jones, L. K., Debernardi, S., Young, B. D., Freemont, P., Verrijzer, C. P., and Saha, V. (2000). Biochemical analyses of the AF10 protein: the extended LAP/PHD-finger mediates oligomerisation. J Mol Biol *299*, 369-378.

Logie, C., and Stewart, A. F. (1995). Ligand-regulated site-specific recombination. Proc Natl Acad Sci U S A *92*, 5940-5944.

Look, A. T. (1997). Oncogenic transcription factors in the human acute leukemias. Science *278*, 1059-1064.

Ma, C., and Staudt, L. M. (1996). LAF-4 encodes a lymphoid nuclear protein with transactivation potential that is homologous to AF-4, the gene fused to MLL in t(4;11) leukemias. Blood *87*, 734-745.

Ma, Q., Alder, H., Nelson, K. K., Chatterjee, D., Gu, Y., Nakamura, T., Canaani, E., Croce, C. M., Siracusa, L. D., and Buchberg, A. M. (1993). Analysis of the murine All-1 gene reveals conserved domains with human ALL-1 and identifies a motif shared with DNA methyltransferases. Proc Natl Acad Sci U S A *90*, 6350-6354.

Mahgoub, N., Parker, R. I., Hosler, M. R., Close, P., Winick, N. J., Masterson, M., Shannon, K. M., and Felix, C. A. (1998). RAS mutations in pediatric leukemias with MLL gene rearrangements. Genes Chromosomes Cancer *21*, 270-275.

Mahillon, J., and Lereclus, D. (1988). Structural and functional analysis of Tn4430: identification of an integrase-like protein involved in the co-integrate-resolution process. Embo J *7*, 1515-1526.

Maki, K., Mitani, K., Yamagata, T., Kurokawa, M., Kanda, Y., Yazaki, Y., and Hirai, H. (1999). Transcriptional inhibition of p53 by the MLL/MEN chimeric protein found in myeloid leukemia. Blood *93*, 3216-3224.

Marschalek, R., Greil, J., Lochner, K., Nilson, I., Siegler, G., Zweckbronner, I., Beck, J. D., and Fey, G. H. (1995). Molecular analysis of the chromosomal breakpoint and fusion transcripts in the acute lymphoblastic SEM cell line with chromosomal translocation t(4;11). Br J Haematol *90*, 308-320.

Messerle, M., Crnkovic, I., Hammerschmidt, W., Ziegler, H., and Koszinowski, U. H. (1997). Cloning and mutagenesis of a herpesvirus genome as an infectious bacterial artificial chromosome. Proc Natl Acad Sci U S A *94*, 14759-14763.

Meyers, E. N., Lewandoski, M., and Martin, G. R. (1998). An Fgf8 mutant allelic series generated by Cre- and Flp-mediated recombination. Nat Genet *18*, 136-141.

Miles, C., Elgar, G., Coles, E., Kleinjan, D. J., van Heyningen, V., and Hastie, N. (1998). Complete sequencing of the Fugu WAGR region from WT1 to PAX6: dramatic compaction and conservation of synteny with human chromosome 11p13. Proc Natl Acad Sci U S A *95*, 13068-13072.

Mitelman, F., Mertens, F., and Johansson, B. (1997). A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. Nat Genet *15 Spec No*, 417-474.

Molnar, A., and Georgopoulos, K. (1994). The Ikaros gene encodes a family of functionally diverse zinc finger DNA-binding proteins. Mol Cell Biol *14*, 8292-8303.

Morgan, B., Sun, L., Avitahl, N., Andrikopoulos, K., Ikeda, T., Gonzales, E., Wu, P., Neben, S., and Georgopoulos, K. (1997). Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros to regulate lymphocyte differentiation. Embo J *16*, 2004-2013.

Morrissey, J., Tkachuk, D. C., Milatovich, A., Francke, U., Link, M., and Cleary, M. L. (1993). A serine/proline-rich protein is fused to HRX in t(4;11) acute leukemias. Blood *81*, 1124-1131.

Morrissey, J. J., Raney, S., and Cleary, M. L. (1997). The FEL (AF-4) protein donates transcriptional activation sequences to Hrx-Fel fusion proteins in leukemias containing T(4;11)(Q21;Q23) chromosomal translocations. Leuk Res *21*, 911-917.

Mountford, P., Zevnik, B., Duwel, A., Nichols, J., Li, M., Dani, C., Robertson, M., Chambers, I., and Smith, A. (1994). Dicistronic targeting constructs: reporters and modifiers of mammalian gene expression. Proc Natl Acad Sci U S A *91*, 4303-4307.

Muyrers, J. P., Zhang, Y., Benes, V., Testa, G., Ansorge, W., and Stewart, A. F. (2000a). Point mutation of bacterial artificial chromosomes by ET recombination. EMBO rep *1*, 239-243.

Muyrers, J. P., Zhang, Y., Buchholz, F., and Stewart, A. F. (2000b). RecE/RecT and Redalpha/Redbeta initiate double-stranded break repair by specifically interacting with their respective partners. Genes Dev *14*, 1971-1982.

Muyrers, J. P., Zhang, Y., Testa, G., and Stewart, A. F. (1999). Rapid modification of bacterial artificial chromosomes by ET-recombination. Nucleic Acids Res *27*, 1555-1557.

Nagy, A., Moens, C., Ivanyi, E., Pawling, J., Gertsenstein, M., Hadjantonakis, A. K., Pirity, M., and Rossant, J. (1998). Dissecting the role of N-myc in development using a single targeting vector to generate a series of alleles. Curr Biol *8*, 661-664.

Nakamura, T., Largaespada, D. A., Lee, M. P., Johnson, L. A., Ohyashiki, K., Toyama, K., Chen, S. J., Willman, C. L., Chen, I. M., Feinberg, A. P., *et al.* (1996a). Fusion of the nucleoporin gene NUP98 to HOXA9 by the chromosome translocation t(7;11)(p15;p15) in human myeloid leukaemia. Nat Genet *12*, 154-158.

Nakamura, T., Largaespada, D. A., Shaughnessy, J. D., Jenkins, N. A., and Copeland, N. G. (1996b). Cooperative activation of Hoxa and Pbx1-related genes in murine myeloid leukaemias. Nat Genet *12*, 149-153.

Nakase, K., Ishimaru, F., Avitahl, N., Dansako, H., Matsuo, K., Fujii, K., Sezaki, N., Nakayama, H., Yano, T., Fukuda, S., *et al.* (2000). Dominant negative isoform of the Ikaros gene in patients with adult B-cell acute lymphoblastic leukemia. Cancer Res *60*, 4062-4065.

Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386-389.

Ng, J., Hart, C. M., Morgan, K., and Simon, J. A. (2000). A Drosophila ESC-E(Z) protein complex is distinct from other polycomb group complexes and contains covalently modified ESC. Mol Cell Biol *20*, 3069-3078.

Nikiforova, M. N., Stringer, J. R., Blough, R., Medvedovic, M., Fagin, J. A., and Nikiforov, Y. E. (2000). Proximity of chromosomal loci that participate in radiation-induced rearrangements in human cells. Science *290*, 138-141.

Nilson, I., Lochner, K., Siegler, G., Greil, J., Beck, J. D., Fey, G. H., and Marschalek, R. (1996). Exon/intron structure of the human ALL-1 (MLL) gene involved in translocations to chromosomal region 11q23 and acute leukaemias. Br J Haematol *93*, 966-972.

Nilson, I., Reichel, M., Ennas, M. G., Greim, R., Knorr, C., Siegler, G., Greil, J., Fey, G. H., and Marschalek, R. (1997). Exon/intron structure of the human AF-4 gene, a member of the AF-4/LAF-4/FMR-2 gene family coding for a nuclear protein with structural alterations in acute leukaemia. Br J Haematol *98*, 157-169.

Okuda, T., Cai, Z., Yang, S., Lenny, N., Lyu, C. J., van Deursen, J. M., Harada, H., and Downing, J. R. (1998). Expression of a knocked-in AML1-ETO leukemia gene inhibits the establishment of normal definitive hematopoiesis and directly generates dysplastic hematopoietic progenitors. Blood *91*, 3134-3143.

Olivero, S., Maroc, C., Beillard, E., Gabert, J., Nietfeld, W., Chabannon, C., and Tonnelle, C. (2000). Detection of different Ikaros isoforms in human leukaemias using real-time quantitative polymerase chain reaction. Br J Haematol *110*, 826-830.

Olson, E. N., Arnold, H. H., Rigby, P. W., and Wold, B. J. (1996). Know your neighbors: three phenotypes in null mutants of the myogenic bHLH gene MRF4. Cell *85*, 1-4.

Orkin, S. H. (2000). Diversification of haematopoietic stem cells to specific lineages. Nat Rev Genet *1*, 57-64.

Orr-Weaver, T. L., and Szostak, J. W. (1983). Yeast recombination: the association between double-strand gap repair and crossing-over. Proc Natl Acad Sci U S A *80*, 4417-4421.

Paro, R., Strutt, H., and Cavalli, G. (1998). Heritable chromatin states induced by the Polycomb and trithorax group genes. Novartis Found Symp *214*, 51-61; discussion 61-56, 104-113.

Pascual, J., Martinez-Yamout, M., Dyson, H. J., and Wright, P. E. (2000). Structure of the PHD zinc finger from human Williams-Beuren syndrome transcription factor. J Mol Biol *304*, 723-729.

Pham, C. T., MacIvor, D. M., Hug, B. A., Heusel, J. W., and Ley, T. J. (1996). Long-range disruption of gene expression by a selectable marker cassette. Proc Natl Acad Sci U S A *93*, 13090-13095.

Pirrotta, V. (1998). Polycombing the genome: PcG, trxG, and chromatin silencing. Cell *93*, 333-336.

Pollock, J. L., Westervelt, P., Kurichety, A. K., Pelicci, P. G., Grisolano, J. L., and Ley, T. J. (1999). A bcr-3 isoform of RARalpha-PML potentiates the development of PML-RARalpha-driven acute promyelocytic leukemia. Proc Natl Acad Sci U S A *96*, 15103-15108.

Prasad, R., Leshkowitz, D., Gu, Y., Alder, H., Nakamura, T., Saito, H., Huebner, K., Berger, R., Croce, C. M., and Canaani, E. (1994). Leucine-zipper dimerization motif encoded by the AF17 gene fused to ALL-1 (MLL) in acute leukemia. Proc Natl Acad Sci U S A *91*, 8107-8111.

Prasad, R., Yano, T., Sorio, C., Nakamura, T., Rallapalli, R., Gu, Y., Leshkowitz, D., Croce, C. M., and Canaani, E. (1995). Domains with transcriptional regulatory activity within the ALL1 and AF4 proteins involved in acute leukemia. Proc Natl Acad Sci U S A *92*, 12160-12164.

Pui, C. H. (2000). Acute lymphoblastic leukemia in children. Curr Opin Oncol *12*, 3-12.

Pui, C. H., Carroll, L. A., Raimondi, S. C., Shuster, J. J., Crist, W. M., and Pullen, D. J. (1994). Childhood acute lymphoblastic leukemia with the t(4;11)(q21;q23): an update. Blood *83*, 2384-2385.

Pui, C. H., Frankel, L. S., Carroll, A. J., Raimondi, S. C., Shuster, J. J., Head, D. R., Crist, W. M., Land, V. J., Pullen, D. J., Steuber, C. P., and et al. (1991). Clinical characteristics and treatment outcome of childhood acute lymphoblastic leukemia with the t(4;11)(q21;q23): a collaborative study of 40 cases. Blood *77*, 440-447.

Rabbitts, T. H. (1994). Chromosomal translocations in human cancer. Nature *372*, 143-149.

Ramirez-Solis, R., Liu, P., and Bradley, A. (1995). Chromosome engineering in mice. Nature *378*, 720-724.

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B. D., Sun, Z. W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C. P., Allis, C. D., and Jenuwein, T. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature *406*, 593-599.

Reeves, R., and Nissen, M. S. (1990). The A.T-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure. J Biol Chem *265*, 8573-8582.

Reichel, M., Gillert, E., Nilson, I., Siegler, G., Greil, J., Fey, G. H., and Marschalek, R. (1998). Fine structure of translocation breakpoints in leukemic blasts with chromosomal translocation t(4;11): the DNA damage-repair model of translocation. Oncogene *17*, 3035-3044.

Renshaw, M. W., McWhirter, J. R., and Wang, J. Y. (1995). The human leukemia oncogene bcr-abl abrogates the anchorage requirement but not the growth factor requirement for proliferation. Mol Cell Biol *15*, 1286-1293.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol *17*, 1030-1032.

Ringrose, L., Chabanis, S., Angrand, P. O., Woodroofe, C., and Stewart, A. F. (1999). Quantitative comparison of DNA looping in vitro and in vivo: chromatin increases effective DNA flexibility at short distances. Embo J *18*, 6630-6641.

Robertson, K. D., Ait-Si-Ali, S., Yokochi, T., Wade, P. A., Jones, P. L., and Wolffe, A. P. (2000). DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters. Nat Genet *25*, 338-342.

Robertson, K. D., and Wolffe, A. P. (2000). DNA methylation in health and disease. Nat Rev Genet *1*, 11-19.

Rodriguez, C. I., Buchholz, F., Galloway, J., Sequerra, R., Kasper, J., Ayala, R., Stewart, A. F., and Dymecki, S. M. (2000). High-efficiency deleter mice show that FLPe is an alternative to Cre-loxP. Nat Genet *25*, 139-140.

Ross, J. A., Potter, J. D., Reaman, G. H., Pendergrass, T. W., and Robison, L. L. (1996). Maternal exposure to potential inhibitors of DNA topoisomerase II and infant leukemia (United States): a report from the Children's Cancer Group. Cancer Causes Control *7*, 581-590.

Rountree, M. R., Bachman, K. E., and Baylin, S. B. (2000). DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. Nat Genet *25*, 269-277.

Roux, S., Terouanne, B., Balaguer, P., Jausons-Loffreda, N., Pons, M., Chambon, P., Gronemeyer, H., and Nicolas, J. C. (1996). Mutation of isoleucine 747 by a threonine alters the ligand responsiveness of the human glucocorticoid receptor. Mol Endocrinol *10*, 1214-1226.

Rowley, J. D. (1998). The critical role of chromosome translocations in human leukemias. Annu Rev Genet *32*, 495-519.

Rowley, J. D., Reshmi, S., Sobulo, O., Musvee, T., Anastasi, J., Raimondi, S., Schneider, N. R., Barredo, J. C., Cantu, E. S., Schlegelberger, B., *et al.* (1997). All patients with the T(11;16)(q23;p13.3) that involves MLL and CBP have treatment-related hematologic disorders. Blood *90*, 535-541.

Rozenblatt-Rosen, O., Rozovskaia, T., Burakov, D., Sedkov, Y., Tillib, S., Blechman, J., Nakamura, T., Croce, C. M., Mazo, A., and Canaani, E. (1998). The C-terminal SET domains of ALL-1 and TRITHORAX interact with the INI1 and SNR1 proteins, components of the SWI/SNF complex. Proc Natl Acad Sci U S A *95*, 4152-4157.

Rozovskaia, T., Rozenblatt-Rosen, O., Sedkov, Y., Burakov, D., Yano, T., Nakamura, T., Petruck, S., Ben-Simchon, L., Croce, C. M., Mazo, A., and Canaani, E. (2000). Self-association of the SET domains of human ALL-1 and of Drosophila TRITHORAX and ASH1 proteins. Oncogene *19*, 351-357.

Sanchis, V., Agaisse, H., Chaufaux, J., and Lereclus, D. (1997). A recombinase-mediated system for elimination of antibiotic resistance gene markers from genetically engineered Bacillus thuringiensis strains. Appl Environ Microbiol *63*, 779-784.

Santulli, B., Kazmierczak, B., Napolitano, R., Caliendo, I., Chiappetta, G., Rippe, V., Bullerdiek, J., and Fusco, A. (2000). A 12q13 translocation involving the HMGI-C gene in richter transformation of a chronic lymphocytic leukemia. Cancer Genet Cytogenet *119*, 70-73.

Satake, N., Ishida, Y., Otoh, Y., Hinohara, S., Kobayashi, H., Sakashita, A., Maseki, N., and Kaneko, Y. (1997). Novel MLL-CBP fusion transcript in therapy-related chronic myelomonocytic leukemia with a t(11;16)(q23;p13) chromosome translocation. Genes Chromosomes Cancer *20*, 60-63.

Schichman, S. A., Caligiuri, M. A., Gu, Y., Strout, M. P., Canaani, E., Bloomfield, C. D., and Croce, C. M. (1994a). ALL-1 partial duplication in acute leukemia. Proc Natl Acad Sci U S A *91*, 6236-6239.

Schichman, S. A., Caligiuri, M. A., Strout, M. P., Carter, S. L., Gu, Y., Canaani, E., Bloomfield, C. D., and Croce, C. M. (1994b). ALL-1 tandem duplication in acute myeloid leukemia with a normal karyotype involves homologous recombination between Alu elements. Cancer Res *54*, 4277-4280.

Schmidt, E. E., Taylor, D. S., Prigge, J. R., Barnett, S., and Capecchi, M. R. (2000). Illegitimate Cre-dependent chromosome rearrangements in transgenic mouse spermatids. Proc Natl Acad Sci U S A *97*, 13702-13707.

Schnabel, C. A., Jacobs, Y., and Cleary, M. L. (2000). HoxA9-mediated immortalization of myeloid progenitors requires functional interactions with TALE cofactors Pbx and Meis. Oncogene *19*, 608-616.

Schultz, D. C., Friedman, J. R., and Rauscher, F. J. (2001). Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2alpha subunit of NuRD. Genes Dev *15*, 428-443.

Schwenk, F., Baron, U., and Rajewsky, K. (1995). A cre-transgenic mouse strain for the ubiquitous deletion of loxP-flanked gene segments including deletion in germ cells. Nucleic Acids Res *23*, 5080-5081.

Schwenk, F., Kuhn, R., Angrand, P. O., Rajewsky, K., and Stewart, A. F. (1998). Temporally and spatially regulated somatic mutagenesis in mice. Nucleic Acids Res *26*, 1427-1432.

Secker-Walker, L. M. (1998). General Report on the European Union Concerted Action Workshop on 11q23, London, UK, May 1997. Leukemia *12*, 776-778.

Seperack, P. K., Strobel, M. C., Corrow, D. J., Jenkins, N. A., and Copeland, N. G. (1988). Somatic and germ-line reverse mutation rates of the retrovirus-induced dilute coat-color mutation of DBA mice. Proc Natl Acad Sci U S A *85*, 189-192.

Shao, Z., Raible, F., Mollaaghababa, R., Guyon, J. R., Wu, C. T., Bender, W., and Kingston, R. E. (1999). Stabilization of chromatin structure by PRC1, a Polycomb complex. Cell *98*, 37-46.

Shashikant, C. S., Carr, J. L., Bhargava, J., Bentley, K. L., and Ruddle, F. H. (1998). Recombinogenic targeting: a new approach to genomic analysis--a review. Gene *223*, 9-20.

Shen, W. F., Rozenfeld, S., Kwong, A., Kom ves, L. G., Lawrence, H. J., and Largman, C. (1999). HOXA9 forms triple complexes with PBX2 and MEIS1 in myeloid cells. Mol Cell Biol *19*, 3051-3061.

Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci U S A *89*, 8794-8797.

Shore, S. K., La Cava, M., Yendapalli, S., and Reddy, E. P. (1994). Structural alterations in the carboxyl-terminal domain of the BCRABL gene product activate its fibroblastic transforming potential. J Biol Chem *269*, 5413-5419.

Simon, J. (1995). Locking in stable states of gene expression: transcriptional control during Drosophila development. Curr Opin Cell Biol *7*, 376-385.

Slany, R. K., Lavau, C., and Cleary, M. L. (1998). The oncogenic capacity of HRX-ENL requires the transcriptional transactivation activity of ENL and the DNA binding motifs of HRX. Mol Cell Biol *18*, 122-129.

Smith, A. J., De Sousa, M. A., Kwabi-Addo, B., Heppell-Parton, A., Impey, H., and Rabbitts, P. (1995). A site-directed chromosomal translocation induced in embryonic stem cells by Cre-loxP recombination. Nat Genet *9*, 376-385.

Sobulo, O. M., Borrow, J., Tomek, R., Reshmi, S., Harden, A., Schlegelberger, B., Housman, D., Doggett, N. A., Rowley, J. D., and Zeleznik-Le, N. J. (1997). MLL is fused to CBP, a histone acetyltransferase, in therapy-related acute myeloid leukemia with a t(11;16)(q23;p13.3). Proc Natl Acad Sci U S A *94*, 8732-8737.

Stassen, M. J., Bailey, D., Nelson, S., Chinwalla, V., and Harte, P. J. (1995). The Drosophila trithorax proteins contain a novel variant of the nuclear receptor type DNA binding domain and an ancient conserved motif found in other chromosomal proteins. Mech Dev *52*, 209-223.

Sternberg, N. L. (1992). Cloning high molecular weight DNA fragments by the bacteriophage P1 system. Trends Genet *8*, 11-16.

Strahl, B. D., and Allis, C. D. (2000). The language of covalent histone modifications. Nature *403*, 41-45.

Strick, R., Strissel, P. L., Borgers, S., Smith, S. L., and Rowley, J. D. (2000). Dietary bioflavonoids induce cleavage in the MLL gene and may contribute to infant leukemia. Proc Natl Acad Sci U S A *97*, 4790-4795.

Strissel, P. L., Espinosa, R., Rowley, J. D., and Swift, H. (1996). Scaffold attachment regions in centromere-associated DNA. Chromosoma *105*, 122-133.

Strissel, P. L., Strick, R., Rowley, J. D., and Zeleznik-Le, N. J. (1998). An in vivo topoisomerase II cleavage site and a DNase I hypersensitive site colocalize near exon 9 in the MLL breakpoint cluster region. Blood *92*, 3793-3803.

Strissel, P. L., Strick, R., Tomek, R. J., Roe, B. A., Rowley, J. D., and Zeleznik-Le, N. J. (2000). DNA structural properties of AF9 are similar to MLL and could act as recombination hot spots resulting in MLL/AF9 translocations and leukemogenesis. Hum Mol Genet *9*, 1671-1679.

Strout, M. P., Marcucci, G., Bloomfield, C. D., and Caligiuri, M. A. (1998). The partial tandem duplication of ALL1 (MLL) is consistently generated by Alu-mediated homologous recombination in acute myeloid leukemia. Proc Natl Acad Sci U S A *95*, 2390-2395.

Sun, L., Crotty, M. L., Sensel, M., Sather, H., Navara, C., Nachman, J., Steinherz, P. G., Gaynon, P. S., Seibel, N., Mao, C., *et al.* (1999a). Expression of dominant-negative Ikaros isoforms in T-cell acute lymphoblastic leukemia. Clin Cancer Res *5*, 2112-2120.

Sun, L., Goodman, P. A., Wood, C. M., Crotty, M. L., Sensel, M., Sather, H., Navara, C., Nachman, J., Steinherz, P. G., Gaynon, P. S., *et al.* (1999b). Expression of aberrantly spliced oncogenic ikaros isoforms in childhood acute lymphoblastic leukemia. J Clin Oncol *17*, 3753-3766.

Sun, L., Heerema, N., Crotty, L., Wu, X., Navara, C., Vassilev, A., Sensel, M., Reaman, G. H., and Uckun, F. M. (1999c). Expression of dominant-negative and mutant isoforms of the antileukemic transcription factor Ikaros in infant acute lymphoblastic leukemia. Proc Natl Acad Sci U S A *96*, 680-685.

Sun, L., Liu, A., and Georgopoulos, K. (1996). Zinc finger-mediated protein interactions modulate Ikaros activity, a molecular control of lymphocyte development. Embo J *15*, 5358-5369.

Swansbury, G. J., Slater, R., Bain, B. J., Moorman, A. V., and Secker-Walker, L. M. (1998). Hematological malignancies with t(9;11)(p21-22;q23)--a laboratory and clinical study of 125 cases. European 11q23 Workshop participants. Leukemia *12*, 792-800.

Taki, T., Kano, H., Taniwaki, M., Sako, M., Yanagisawa, M., and Hayashi, Y. (1999). AF5q31, a newly identified AF4-related gene, is fused to MLL in infant acute

lymphoblastic leukemia with ins(5;11)(q31;q13q23). Proc Natl Acad Sci U S A *96*, 14535-14540.

Taki, T., Sako, M., Tsuchida, M., and Hayashi, Y. (1997). The t(11;16)(q23;p13) translocation in myelodysplastic syndrome fuses the MLL gene to the CBP gene. Blood *89*, 3945-3950.

Tanabe, S., Bohlander, S. K., Vignon, C. V., Espinosa, R., Zhao, N., Strissel, P. L., Zeleznik-Le, N. J., and Rowley, J. D. (1996). AF10 is split by MLL and HEAB, a human homolog to a putative Caenorhabditis elegans ATP/GTP-binding protein in an invins(10;11)(p12;q23q12). Blood *88*, 3535-3545.

Tang, A. H., Neufeld, T. P., Rubin, G. M., and Muller, H. A. (2001). Transcriptional regulation of cytoskeletal functions and segmentation by a novel maternal pair-rule gene, lilliputian. Development *128*, 801-813.

Thomas, K. R., and Capecchi, M. R. (1986). Introduction of homologous DNA sequences into mammalian cells induces mutations in the cognate gene. Nature *324*, 34-38.

Thomas, K. R., and Capecchi, M. R. (1987). Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. Cell *51*, 503-512.

Thomas, K. R., Deng, C., and Capecchi, M. R. (1992). High-fidelity gene targeting in embryonic stem cells by using sequence replacement vectors. Mol Cell Biol *12*, 2919-2923.

Thorsteinsdottir, U., Sauvageau, G., Hough, M. R., Dragowska, W., Lansdorp, P. M., Lawrence, H. J., Largman, C., and Humphries, R. K. (1997). Overexpression of HOXA10 in murine hematopoietic cells perturbs both myeloid and lymphoid differentiation and leads to acute myeloid leukemia. Mol Cell Biol *17*, 495-505.

Tillib, S., Petruk, S., Sedkov, Y., Kuzin, A., Fujioka, M., Goto, T., and Mazo, A. (1999). Trithorax- and Polycomb-group response elements within an Ultrabithorax transcription maintenance unit consist of closely situated but separable sequences. Mol Cell Biol *19*, 5189-5202.

Tillib, S., Sedkov, Y., Mizrokhi, L., and Mazo, A. (1995). Conservation of structure and expression of the trithorax gene between Drosophila virilis and Drosophila melanogaster. Mech Dev *53*, 113-122.

Tkachuk, D. C., Kohler, S., and Cleary, M. L. (1992). Involvement of a homolog of Drosophila trithorax by 11q23 chromosomal translocations in acute leukemias. Cell *71*, 691-700.

Uckun, F. M., Herman-Hatten, K., Crotty, M. L., Sensel, M. G., Sather, H. N., Tuel-Ahlgren, L., Sarquis, M. B., Bostrom, B., Nachman, J. B., Steinherz, P. G., *et al.* (1998). Clinical significance of MLL-AF4 fusion transcript expression in the absence of a cytogenetically detectable t(4;11)(q21;q23) chromosomal translocation. Blood *92*, 810-821.

Van Deursen, J., Fornerod, M., Van Rees, B., and Grosveld, G. (1995). Cre-mediated site-specific translocation between nonhomologous mouse chromosomes. Proc Natl Acad Sci U S A *92*, 7376-7380.

van Lohuizen, M., Tijms, M., Voncken, J. W., Schumacher, A., Magnuson, T., and Wientjens, E. (1998). Interaction of mouse polycomb-group (Pc-G) proteins Enx1 and Enx2 with Eed: indication for separate Pc-G complexes. Mol Cell Biol *18*, 3572-3579.

Versteege, I., Sevenet, N., Lange, J., Rousseau-Merck, M. F., Ambros, P., Handgretinger, R., Aurias, A., and Delattre, O. (1998). Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. Nature *394*, 203-206.

Vidal, M. (2001). A biological atlas of functional maps. Cell *104*, 333-339.

von Lindern, M., van Baal, S., Wiegant, J., Raap, A., Hagemeijer, A., and Grosveld, G. (1992). Can, a putative oncogene associated with myeloid leukemogenesis, may be activated by fusion of its 3' half to different genes: characterization of the set gene. Mol Cell Biol *12*, 3346-3355.

Wang, J. H., Nichogiannopoulou, A., Wu, L., Sun, L., Sharpe, A. H., Bigby, M., and Georgopoulos, K. (1996). Selective defects in the development of the fetal and adult lymphoid system in mice with an Ikaros null mutation. Immunity *5*, 537-549.

Westervelt, P., and Ley, T. J. (1999). Seed versus soil: the importance of the target cell for transgenic models of human leukemias. Blood *93*, 2143-2148.

Wiedemann, L. M., MacGregor, A., and Caldas, C. (1999). Analysis of the region of the 5' end of the MLL gene involved in genomic duplication events. Br J Haematol *105*, 256-264.

Winandy, S., Wu, P., and Georgopoulos, K. (1995). A dominant mutation in the Ikaros gene leads to rapid development of leukemia and lymphoma. Cell *83*, 289-299.

Wittwer, F., van der Straten, A., Keleman, K., Dickson, B. J., and Hafen, E. (2001). Lilliputian: an AF4/FMR2-related protein that controls cell identity and cell growth. Development *128*, 791-800.

Woodage, T., Basrai, M. A., Baxevanis, A. D., Hieter, P., and Collins, F. S. (1997). Characterization of the CHD family of proteins. Proc Natl Acad Sci U S A *94*, 11472-11477.

Yagi, H., Deguchi, K., Aono, A., Tani, Y., Kishimoto, T., and Komori, T. (1998). Growth disturbance in fetal liver hematopoiesis of Mll-mutant mice. Blood *92*, 108-117.

Yang, X. W., Model, P., and Heintz, N. (1997). Homologous recombination based modification in Escherichia coli and germline transmission in transgenic mice of a bacterial artificial chromosome. Nat Biotechnol *15*, 859-865.

Yano, T., Nakamura, T., Blechman, J., Sorio, C., Dang, C. V., Geiger, B., and Canaani, E. (1997). Nuclear punctate distribution of ALL-1 is conferred by distinct elements at the N terminus of the protein. Proc Natl Acad Sci U S A *94*, 7286-7291.

Yergeau, D. A., Hetherington, C. J., Wang, Q., Zhang, P., Sharpe, A. H., Binder, M., Marin-Padilla, M., Tenen, D. G., Speck, N. A., and Zhang, D. E. (1997). Embryonic lethality and impairment of haematopoiesis in mice heterozygous for an AML1-ETO fusion gene. Nat Genet *15*, 303-306.

Yu, B. D., Hanson, R. D., Hess, J. L., Horning, S. E., and Korsmeyer, S. J. (1998). MLL, a mammalian trithorax-group gene, functions as a transcriptional maintenance factor in morphogenesis. Proc Natl Acad Sci U S A *95*, 10632-10636.

Yu, B. D., Hess, J. L., Horning, S. E., Brown, G. A., and Korsmeyer, S. J. (1995). Altered Hox expression and segmental identity in Mll-mutant mice. Nature *378*, 505-508.

Yuge, M., Nagai, H., Uchida, T., Murate, T., Hayashi, Y., Hotta, T., Saito, H., and Kinoshita, T. (2000). HSNF5/INI1 gene mutations in lymphoid malignancy. Cancer Genet Cytogenet *122*, 37-42.

Zeleznik-Le, N. J., Harden, A. M., and Rowley, J. D. (1994). 11q23 translocations split the "AT-hook" cruciform DNA-binding region and the transcriptional repression domain from the activation domain of the mixed-lineage leukemia (MLL) gene. Proc Natl Acad Sci U S A *91*, 10610-10614.

Zhang, Y., Buchholz, F., Muyrers, J. P., and Stewart, A. F. (1998a). A new logic for DNA engineering using recombination in Escherichia coli. Nat Genet *20*, 123-128.

Zhang, Y., LeRoy, G., Seelig, H. P., Lane, W. S., and Reinberg, D. (1998b). The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. Cell *95*, 279-289.

Zhang, Y., Muyrers, J. P., Testa, G., and Stewart, A. F. (2000). DNA cloning by homologous recombination in Escherichia coli. Nat Biotechnol *18*, 1314-1317.

Zhang, Y., Riesterer, C., Ayrall, A. M., Sablitzky, F., Littlewood, T. D., and Reth, M. (1996). Inducible site-directed recombination in mouse embryonic stem cells. Nucleic Acids Res *24*, 543-548.

Zheng, B., Sage, M., Sheppeard, E. A., Jurecic, V., and Bradley, A. (2000). Engineering mouse chromosomes with Cre-loxP: range, efficiency, and somatic applications. Mol Cell Biol *20*, 648-655.

Ziemin-van der Poel, S., McCabe, N. R., Gill, H. J., Espinosa, R., Patel, Y., Harden, A., Rubinelli, P., Smith, S. D., LeBeau, M. M., Rowley, J. D., and et al. (1991). Identification of a gene, MLL, that spans the breakpoint in 11q23 translocations associated with human leukemias. Proc Natl Acad Sci U S A *88*, 10735-10739.

Zimonjic, D. B., Pollock, J. L., Westervelt, P., Popescu, N. C., and Ley, T. J. (2000). Acquired, nonrandom chromosomal abnormalities associated with the development of acute promyelocytic leukemia in transgenic mice. Proc Natl Acad Sci U S A *97*, 13306-13311.

Zink, D., Cremer, T., Saffrich, R., Fischer, R., Trendelenburg, M. F., Ansorge, W., and Stelzer, E. H. (1998). Structure and dynamics of human interphase chromosome territories in vivo. Hum Genet *102*, 241-251.

Acknowledgements

I wish to thank first of all Dr. Francis Stewart, my supervisor, for his enthusiasm, guidance and continuous support. I am grateful to Dr. Iain Mattaj, Dr. Stephen Cohen and Dr. Renato Paro of my EMBL thesis committee for their insightful comments and suggestions throughout the development of this project.

A great thanks to Dr. Terry Rabbitts, for his advice and support, and for an exciting collaboration.

Thanks to Vladimir Benes of the EMBL sequencing facility for a great effort with sequencing all the BACs of this project and for his constant motivation.

My deep gratitude to Kristina Vintersten of the EMBL transgenic core facility, for the generation of all the mice described in this thesis and for her contagious excitement.

Thanks to the staff of the EMBL animal house for their support.

Thanks to all members of the Stewart lab, past and present, for creating a great atmosphere in the lab. I am especially thankful to Frank van der Hoeven and Konstantinos Anastassiadis for always being there to answer my questions, and to Yu Min Zhang and Joep Muyrers for their great advice on ET recombination.

Many thanks to the colleagues from the Centre of Genome Research in Edinburgh, especially Austin Smith, Andrew Smith, Ian Chambers and Dani Nebeniun-Oosthuizen. Thanks for sharing with me work and enthusiasm in the final phase of the Mll project.

I am grateful to Meinrad Busslinger for the Ikaros BACs.

A great many thanks to you, Francis and Michelle, for your commitment to reading.

I treasure the memories of these years with all the friends at EMBL: Davide and Stefano for a walk in the snow and much more; Michi and the whole crowd wich dared to go from genes to thoughts.

Thanks to the staff of the photolab, for their great help, and to the EMBL librarians for their wonderful comittment.

And thank you, mom, for always smiling when I built improbable airplanes with wooden bricks.